Submitted to
manuscript

Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity

Sébastien Gadat
Toulouse School of Economics, Université Toulouse 1 Capitole sebastien.gadat@math.univ-toulouse.fr,
http://perso.math.univ-toulouse.fr/gadat/

Fabien Panloup
Laboratoire Angevin de Recherche en Mathématiques, Université d'Angers fabien.panloup@math.univ-angers.fr,
http://blog.univ-angers.fr/panloup/

This paper is devoted to the non-asymptotic control of the mean-squared error for the Ruppert-Polyak stochastic averaged gradient descent introduced in the seminal contributions of [24] and [26]. In our main results, we establish non-asymptotic tight bounds (optimal with respect to the Cramer-Rao lower bound) in a very general framework that includes the uniformly strongly convex case as well as the one where the function f to be minimized satisfies a weaker Kurdyka-Łojiasewicz-type condition [18, 19]. In particular, it makes it possible to recover some pathological examples such as on-line learning for logistic regression (see [2]) and recursive quantile estimation (an even non-convex situation). Finally, our bound is optimal when the decreasing step $(\gamma_n)_{n\geq 1}$ satisfies: $\gamma_n = \gamma n^{-\beta}$ with $\beta = 3/4$, leading to a second-order term in $O(n^{-5/4})$.

Key words: stochastic algorithms, optimization, averaging
MSC2000 subject classification: 62L20,80M50,68W25
OR/MS subject classification: Primary: Stochastic optimization algorithm

1. Introduction

1.1. Averaging principle for stochastic algorithms Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that belongs to $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$, *i.e.*, the space of twice differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$ with continuous second partial derivatives. Let us assume that ∇f admits the following representation: a measurable function $\Lambda : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ and a random variable Z with values in $\mathbb{R}^p$ exist such that Z is distributed according to μ such that:

$$\forall \theta \in \mathbb{R}^d, \quad \nabla f(\theta) = \mathbb{E}[\Lambda(\theta, Z)]. \tag{1}$$

In this case, the Robbins-Monro procedure introduced in the seminal contribution [25] is built with an i.i.d. sequence of observations $(Z_i)_{i\geq 1}$ distributed according to μ. It is well known that under appropriate assumptions, the minimizers of f can be approximated through the recursive stochastic algorithm $(\theta_n)_{n\geq 0}$ defined by: $\theta_0 \in \mathbb{R}^d$ and

$$\forall n \geq 0, \quad \theta_{n+1} = \theta_n - \gamma_{n+1}\Lambda(\theta_n, Z_{n+1}), \tag{2}$$

where $(\gamma_n)_{n\geq 1}$ denotes a non-increasing sequence of positive numbers such that:

$$\Gamma_n := \sum_{k=1}^{n} \gamma_k \xrightarrow{n\longrightarrow+\infty} +\infty \quad \text{and} \quad \gamma_n \xrightarrow{n\longrightarrow+\infty} 0.$$

Equation (2) is sometimes written as a noisy gradient descent:

$$\forall n \geq 0, \quad \theta_{n+1} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta M_{n+1}, \tag{3}$$

where $(\Delta M_{n+1})_{n\geq 0}$ stands for the sequence of noises defined by:

$$\forall n \geq 0, \quad \Delta M_{n+1} = -\Lambda(\theta_n, Z_{n+1}) + \nabla f(\theta_n). \tag{4}$$

Equation (1) shows that the sequence $(\Delta M_n)_{n\geq 1}$ is a sequence of martingale increments, *i.e.*

$$\forall n \geq 1, \quad \mathbb{E}\left[\Delta M_{n+1} | \mathcal{F}_n\right] = 0,$$

where $(\mathcal{F}_n)_{n\geq 0}$ is the filtration defined by $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$ for $n \geq 1$, $\mathcal{F}_0$ is the trivial σ-field and for a given σ-field $\mathcal{G}$, $\mathbb{E}[\,.\,|\mathcal{G}]$ stands for the related conditional expectation.

The standard averaging procedure of Ruppert-Polyak (referred to as RP averaging) consists in introducing a Cesaro average over the iterations of the Robbins-Monro sequence defined by:

$$\hat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n} \theta_k, \quad n \geq 1.$$

It is well known that such an averaging procedure is a way to improve the convergence properties of the original algorithm $(\theta_n)_{n\geq 1}$ by minimizing the asymptotic variance induced by the algorithm. More precisely, when f is a strongly convex function and possesses a unique minimum $\theta^\star$, $\sqrt{n}(\hat{\theta}_n - \theta^\star)_{n\geq 1}$ converges in distribution to a Gaussian law whose variance attains the Cramer-Rao lower bound of any unbiased estimation of $\theta^\star$ (see Theorem 1 for a precise statement of this state of the art result).

Such results are usually achieved *asymptotically* in general situations where f is assumed to be strongly convex, we refer to [24] for the initial asymptotic description and to [14] for some more general results. In [3], a non-asymptotic optimal (with a sharp first order term) result is obtained in the strongly convex situation under restrictive moment assumptions on the noisy gradients. The problem is also tackled non asymptotically in some specific cases when the strong convexity property fails (on-line logistic regression [2], recursive median estimation [10, 9] for example). Nevertheless, a general result for strongly convex or not situations under some mild conditions on the noise while preserving a sharp optimal $O(n^{-1})$ rate of convergence of the $\mathbb{L}^2$-risk is yet missing.

In this paper, our objective is to derive optimal non-asymptotic $\mathbb{L}^2$-risks bounds for the Ruppert-Polyak (RP) algorithm under some very general assumptions beyond the traditional convexity point of view. This goal is achieved in two steps. In a first stage, we obtain a general theorem from a sharp study of the RP-dynamics under a so-called *consistency* assumption on the original procedure $(\theta_n)_{n\geq 0}$ (see Section 2.3). In this result, the bound is optimal at the first order since it attains the Cramer-Rao bound (*i.e.* rate in $O(n^{-1})$ with the lowest variance) and provides a second order term which is better than similar results of the literature (see Table 1 for details). In a second stage, we show that our consistency assumption holds in the strongly convex case but also under the so-called *Kurdyka- Lojasiewicz inequality* (see [18, 19]), which is a much weaker situation than the strongly convex settings. This second part leads to some considerable improvements of state of the art result since important applications are not tackled by the strongly convex setting, such as the on-line logistic regression example and the recursive quantile approximation.

1.2. Polyak-Juditsky central limit theorem To assess the quality of a non-asymptotic control of the sequences $(\hat{\theta}_n)_{n\geq 0}$, we recall the CLT associated with $(\hat{\theta}_n)_{n\geq 0}$, whose statement is adapted from [24][1] with the strongly convex assumption ($\mathbf{H_{SC(\alpha)}}$):

[1] In [24], the result is stated in a slightly more general framework with the help of a Lyapunov function. We have chosen to simplify the statement for the sake of readability.

Assumption $\mathbf{H_{SC(\alpha)}}$ - Strongly convex function *f is a strongly convex function of parameter $\alpha > 0$ in the set:*

$$SC(\alpha) := \left\{ f \in \mathcal{C}^2(\mathbb{R}^d) : D^2 f - \alpha I_d \geq 0 \right\} \tag{5}$$

where $D^2 f$ stands for the Hessian matrix of f and inequality $A \geq 0$ for any matrix A has to be understood in the sense of quadratic forms.

The set $SC(\alpha)$ captures many practical situations such as the least square optimization problem in statistical linear models for example.

Theorem 1 ([24]) *Assume that:*
i) the function f is in $SC(\alpha)$ and $x \longmapsto D^2 f(x)$ is bounded.
ii) $\gamma_n \xrightarrow{n \to +\infty} 0$ and $\gamma_n^{-1}(\gamma_n - \gamma_{n+1}) = o_{+\infty}(\gamma_n)$,
iii) the convergence in probability of the conditional covariance holds, i.e.,

$$\lim_{n \longrightarrow +\infty} \mathbb{E}[\Delta M_{n+1} \Delta M_{n+1}^T | \mathcal{F}_n] = S^\star,$$

then

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) \xrightarrow[n \longrightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma^\star),$$

where

$$\Sigma^\star = \{D^2 f(\theta^\star)\}^{-1} S^\star \{D^2 f(\theta^\star)\}^{-1}. \tag{6}$$

Theorem 1 shows that the Ruppert-Polyak averaging produces an asymptotically optimal algorithm whose rate of convergence is $O(n^{-1})$, which is minimax optimal in the class of strongly convex stochastic minimization problems (see, *e.g.* [21]). Moreover, the asymptotic variance is also optimal because it attains the Cramer-Rao lower bound (see, *e.g.* [24, 11]).

It is also important to observe that $(\hat{\theta}_n)_{n \geq 0}$ is an adaptive sequence since the previous result is obtained independently of the size of $D^2 f(\theta^\star)$ as soon as the sequence $(\gamma_n)_{n \geq 1}$ is chosen as $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0, 1)$.

2. Main results

2.1. Summary of our contributions As pointed out by many authors in some recent works (we refer to [2], [3] and [10], among others), even though very general, Theorem 1 has the usual drawback of being only asymptotic with respect to n. To bypass this weak quantitative result, some improvements are then obtained for various particular cases of minimization problems (*e.g.*, logistic regression, least square minimization, median and quantile estimations) in the literature.

Below, we are interested in deriving some non-asymptotic inequality results on the RP averaging for the minimization of f. More precisely, in our main results, we will discuss on L^2-risk bounds of the following form:

$$\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] \leq c_1 n^{-1} + c_2 n^{-\rho} \quad \text{with } \rho > 1.$$

We will say that the first-order term in this non-asymptotic bound attains the Cramer-Rao lower-bound if $c_1 = \mathrm{Tr}(\Sigma^\star)$. Such a bound is stated in a general setting in Theorem 8 and then applied in a series of contexts. In order to give a roadmap of these applications, we provide a summary of our contributions in the following table, enriched with a comparison with the existing results in the literature:

	Setting	Cramer-Rao	2nd order v_n	$\gamma_n = \gamma_1 n^{-\beta}$	Anytime
Our work	Strong. Convex Convex (Smooth KL) Logist. Reg. (KL) Recurs. Quantile (KL)	Yes : $\frac{Tr(\Sigma^\star)}{n}$	$n^{-(\beta+\frac{1}{2})\wedge(2-\beta)}$, $v_n^\star = O(n^{-\frac{5}{4}})$	$\beta \in (1/2,1)$ $\beta^\star = 3/4$	Yes
BM(11) [3]	Strong. Convex	Yes : $\frac{Tr(\Sigma^\star)}{n}$	$n^{-(\beta+\frac{1}{2})\wedge(\frac{3}{2}-\beta)}$, $v_n^\star = O(n^{-\frac{7}{6}})$	$\beta \in (1/2,1)$ $\beta^\star = 2/3$	Yes
BM(11) [3]	Convex Logist. Reg. Recurs. Quantile	No: $O(n^{-1/2})$ No: $O(n^{-1/2})$ $\emptyset$	$\emptyset$	$\beta = 1/2$	Yes
B(14) [2]	Logist. Reg.	No: $O\left(\frac{1}{n\lambda^2_{min}\{D^2 f(\theta^\star)\}}\right)$	$\emptyset$	$\beta = 1/2$	No
CCGB(17) [9]	Recurs. Quantile	No: $O\left(\frac{1}{n}\right)$	$n^{-(\beta+\frac{1}{2})\wedge(\frac{3}{2}-\beta)}$, $v_n^\star = O(n^{-\frac{7}{6}})$	$\beta \in (1/2,1)$ $\beta^\star = 2/3$	Yes

TABLE 1. Overview of our results and comparisons with the literature. $v_n^\star$ refers to the optimal (smallest) size of the second-order term when β is chosen equal to $\beta^\star$.

2.2. Notations For any vector $y \in \mathbb{R}^d$, y^T denotes the transpose of y, whereas $|y|$ is the Euclidean norm of y in $\mathbb{R}^d$. The set $\mathcal{M}_d(\mathbb{R})$ refers to the set of squared real matrices of size $d \times d$ and the tensor product $\otimes 2$ is used to refer to the following quadratic form:

$$\forall M \in \mathcal{M}_d(\mathbb{R}) \quad \forall y \in \mathbb{R}^d \qquad My^{\otimes 2} = y^T M y.$$

I_d is the identity matrix in $\mathcal{M}_d(\mathbb{R})$ and $\mathcal{O}_d(\mathbb{R})$ denotes the set of orthonormal real matrices of size $d \times d$:

$$\mathcal{O}_d(\mathbb{R}) := \left\{ Q \in \mathcal{M}_d(\mathbb{R}) : Q^T Q = I_d \right\}.$$

Finally, the notation $\|\,.\,\|$ corresponds to a (non-specified) norm on $\mathcal{M}_d(\mathbb{R})$.

For two positive sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, the notation $a_n \lesssim b_n$ refers to the domination relationship, *i.e.* $a_n \leq c\, b_n$ where $c > 0$ is independent of n. The binary relationship $a_n = \mathcal{O}(b_n)$ then holds if and only if $|a_n| \lesssim |b_n|$. Finally, if for all $n \in \mathbb{N}$, $b_n \neq 0$, $a_n = o(b_n)$ if $\lim \frac{a_n}{b_n} = 0$ when $n \longrightarrow +\infty$.

In the rest of the paper, we assume that f satisfies the following properties:

$$\lim_{|x|\to+\infty} f(x) = +\infty \quad \text{and} \quad \{x \in \mathbb{R}^d, \nabla f(x) = 0\} = \{\theta^\star\}, \tag{7}$$

where $\theta^\star$ is thus the unique minimum of f. Without loss of generality, we also assume that $f(\theta^\star) = 0$.

We also consider the common choice for $(\gamma_n)_{n\geq 1}$ (for $\gamma > 0$ and $\beta \in (0,1)$):

$$\forall n \geq 1 \qquad \gamma_n = \gamma n^{-\beta}.$$

In particular, we have $\Gamma_n \sim \frac{\gamma}{1-\beta} n^{1-\beta} \longrightarrow +\infty$ and $\gamma_n \longrightarrow 0$ as $n \longrightarrow +\infty$.

The rest of this section is devoted to the statement of our main results. In Subsection 2.3, we state our main general result (Theorem 2) under some general assumptions on the noise part and on the behavior of the L^p-norm of the **original procedure** $(\theta_n)_{n\geq 1}$ ($(L^p, \sqrt{\gamma_n})$-*consistency*). Then, in the next subsections, we provide some settings where this consistency condition is satisfied: under a strong convexity assumption in Subsection 2.4.1 and under a weaker Kurdyka-Łojiasewicz-type assumption in Subsection 2.4.2.

2.3. Non asymptotic adaptive and optimal inequality Our first main result is Theorem 2 and we introduce the following definition.

Definition 2.1 ($(L^p,\sqrt{\gamma_n})$-consistency) *Let $p>0$. We say that a sequence $(\theta_n)_{n\geq 1}$ satisfies the $(L^p,\sqrt{\gamma_n})$-consistency (convergence rate condition) if $\left(\frac{\theta_n}{\sqrt{\gamma_n}}\right)_{n\geq 1}$ is bounded in L^p, i.e., if:*

$$\exists\, c_p>0 \quad \forall n\geq 1 \qquad \mathbb{E}|\theta_n|^p \leq c_p\{\gamma_n\}^{\frac{p}{2}}.$$

Note that according to the Jensen inequality, the $(L^p,\sqrt{\gamma_n})$-consistency implies the $(L^q,\sqrt{\gamma_n})$-consistency for any $0<q<p$. As mentioned before, this definition refers to the behaviour of the crude procedure $(\theta_n)_{n\geq 1}$ defined by Equation (2). We will prove that Definition 2.1 is a key property to derive sharp non-asymptotic bounds for the RP-averaged algorithm $(\hat{\theta}_n)_{n\geq 1}$ (see Theorem 2 below).

We also introduce a smoothness assumption on the covariance of the martingale increment:

Assumption ($\mathbf{H_S}$) - Covariance of the martingale increment *The covariance of the martingale increment introduced in (4) satisfies:*

$$\mathbb{E}\left[\Delta M_{n+1}\Delta M_{n+1}^t|\mathcal{F}_n\right] = S(\theta_n) \qquad a.s.$$

where $S:\mathbb{R}^d\to\mathcal{M}_d(\mathbb{R})$ is a Lipschitz continuous function:

$$\exists L>0 \quad \forall(\theta_1,\theta_2)\in\mathbb{R}^d \qquad \|S(\theta_1)-S(\theta_2)\|\leq L|\theta_1-\theta_2|.$$

When compared to Theorem 1 *iii*), Assumption ($\mathbf{H_S}$) is more restrictive but in fact corresponds to the usual framework. Under additional technicalities, this assumption may be relaxed to a local Lipschitz behaviour of S. For reasons of clarity, we preferred to reduce our purpose to this reasonable setting. We now state our main general result:

Theorem 2 (Optimal non-asymptotic bound for the averaging procedure) *Let $\gamma_n=\gamma n^{-\beta}$ with $\beta\in(1/2,1)$. Assume that $(\theta_n)_{n\geq 1}$ is $(L^4,\sqrt{\gamma_n})$-consistent and that Assumption ($\mathbf{H}_S$) holds. Suppose moreover that $D^2f(\theta^\star)$ is positive-definite.Then, a large enough C exists such that:*

$$\forall n\in\mathbb{N}^\star \qquad \mathbb{E}\left[|\hat{\theta}_n-\theta^\star|^2\right]\leq \frac{\mathrm{Tr}(\Sigma^\star)}{n}+Cn^{-r_\beta}, \tag{8}$$

where $\Sigma^\star$ is defined in Equation (6) (with $S^\star=S(\theta^\star)$) and

$$r_\beta=\left(\beta+\frac{1}{2}\right)\wedge(2-\beta).$$

In particular, $r_\beta>1$ for all $\beta\in(1/2,1)$ and $\beta\longmapsto r_\beta$ attains its maximum for $\beta=3/4$, which yields:

$$\forall n\in\mathbb{N}^\star \qquad \mathbb{E}\left[|\hat{\theta}_n-\theta^\star|^2\right]\leq \frac{\mathrm{Tr}(\Sigma^\star)}{n}+Cn^{-5/4}.$$

The result stated by Theorem 2 deserves several remarks.

• *Sharpness of the L^2-bound/First and second order terms*: as mentioned before, we obtain the exact optimal rate $O(n^{-1})$ with the sharp constant $\mathrm{Tr}(\Sigma^\star)$ as shown by Theorem 1. Hence, at the first order, Theorem 2 shows that the averaging procedure is minimax optimal with respect to

the Cramer-Rao lower bound. Moreover, the result is adaptive with respect to the value of the Hessian $D^2f(\theta^\star)$: any sequence $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (1/2,1)$ and $\gamma > 0$, regardless the value of β or γ, produces the result of Theorem 2. We should note that such an adaptive property does not hold for the initial sequence $(\theta_n)_{n\geq 1}$ as proved by the central limit theorem satisfied by $(\theta_n)_{n\geq 1}$ (see [13] for example).

Even though any value of $\beta \in (1/2,1)$ yields a $\frac{\mathrm{Tr}(\Sigma^\star)}{n}$ leading term, the "optimal" choice of β remains unclear. In [3] and [9], $\beta = 2/3$ is motivated by the optimization of the second order term. In particular, [3] obtains in the strongly convex case an upper bound of the order $\frac{Tr(\Sigma^\star)}{n} + O(n^{-7/6})$. In our work, Theorem 2 also improves this second order term since the choice $\beta = 3/4$ leads to an upper bound of the order $\frac{Tr(\Sigma^\star)}{n} + O(n^{-5/4})$. Moreover, for any value of $\beta \in (1/2,1)$, the second order term in Theorem 2 is $O(n^{-(\beta+1/2)\wedge(2-\beta)})$, which is always better than $O(n^{-(\beta+1/2)\wedge(3/2-\beta)})$, the one of [3]. For further comments on this topic (including the particular case of null third derivatives), we refer to Section 3.2.

• *Idea of the proof/Assumptions*: the proof of Theorem 2 is achieved through a spectral analysis of the second-order Markov chain induced by $(\hat{\theta}_n)_{n\geq 1}$. This spectral analysis requires a preliminary linearization step of the drift from $\hat{\theta}_n$ to $\hat{\theta}_{n+1}$. The cost of this linearization is absorbed by a preliminary control of the initial sequence $(\theta_n)_{n\geq 1}$, obtained with the $(L^p, \sqrt{\gamma_n})$-consistency for $p = 4$ (see Proposition 2.1 and Theorem 4 for results on the $(L^p,\sqrt{\gamma_n})$-consistency). Let us remark that this linearization approach implies that our result *a priori* applies regardless the global assumptions on the objective function: we only impose a local curvature around $\theta^\star$.

• *Anytime strategy*: an important feature of on-line optimization algorithm is the *anytime property*, *i.e.*, the ability of the algorithm to produce an optimal performance regardless the choice of the stopping iteration time: in commonly encountered situations, the final number of iterations is not known in advance. In general, a such anytime property fails when the step-size sequence depends on the final number of iterations. One common way to bypass this issue is to use the doubling trick strategy (see, *e.g.* [12]) , which produces an anytime algorithm and that degrades the final rate with an almost negligible multiplicative logarithmic term. However, for the Ruppert-Polyak algorithm, the consequence of such a doubling trick on the initial SGD sequence remains unclear for $(\hat{\theta}_n)_{n\geq 1}$ because $(\hat{\theta}_n)_{n\geq 1}$ uses all the iterates of the SGD sequence with uniform weights.

As indicated in our Theorem 2, in [3] (for strongly convex function) and [9] (quantile estimation), the sequence $(\gamma_n)_{n\geq 1}$ is chosen independently of the final horizon time, and the procedure is therefore anytime. Oppositely, the step-size sequence proposed in [2] highly depends on the final number of iteration (the proposed sequence is constant and equal to $\frac{1}{2R^2\sqrt{n}}$ where n is the stopping time, so that the method of [2] for on-line logistic regression is not anytime).

2.4. $(L^p,\sqrt{\gamma_n})$-consistency As indicated in Theorem 2, the control of the moments of the sequence $(\theta_n - \theta^\star)_{n\geq 1}$ is an important ingredient to derive the optimal bound (8). We first present how to obtain a such moment upper bound in the standard strongly convex case, and then in some more general cases without convexity.

2.4.1. $(L^p,\sqrt{\gamma_n})$-consistency with strong convexity In this section, we temporarily restrict our study to the classical setting $\mathbf{H_{SC(\alpha)}}$ and we need to add an additional condition on the noise, denoted by $(\mathbf{H^{SC}_{\Sigma_p}})$:

Assumption $(\mathbf{H^{SC}_{\Sigma_p}})$ - Moments of the martingale increment *For a given $p \in \mathbb{N}^\star$, the sequence of martingale increments satisfies: a constant Σ_p exists such that for any $n \in \mathbb{N}$:*

$$\mathbb{E}[|\Delta M_{n+1}|^{2p}|\mathcal{F}_n] \leq \Sigma_p(1 + (f(\theta_n))^p \qquad a.s.$$

We emphasize that even though Assumption $\mathbf{H_{SC(\alpha)}}$ is a potentially restrictive assumption on f, the one on the martingale increments is not restrictive and allows a polynomial dependency in $f(\theta_n)$ of the moments of ΔM_n, which is much weaker than the one used in Theorem 3 of [3]. For example, such an assumption holds in the case of the recursive linear least square problem. In that case, we retrieve the baseline assumption introduced in [13] that only provides an almost sure convergence of $(\theta_n)_{n\geq 1}$ towards $\theta^\star$ without any rate. In this setting, we can state the following proposition, whose proof is left to the reader and up to some minor modifications, is contained in the more general result stated in Theorem 6 (see Section 2.4.2).

Proposition 2.1 *Assume that a $\alpha > 0$ exists such that f is $\mathbf{H_{SC(\alpha)}}$ and that $x \longmapsto D^2 f(x)$ is Lipschitz bounded. If the sequence $(\Delta M_n)_{n\geq 1}$ satisfies $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$, then $(\theta_n)_{n\geq 1}$ is $(L^p,\sqrt{\gamma_n})$-consistent for any $p \geq 1$:*

$$\forall p \geq 1 \quad \exists C_p > 0 \qquad \mathbb{E}|\theta_n - \theta^\star|^p \leq C_p \{\gamma_n\}^{p/2}.$$

An immediate consequence of Proposition 2.1 and of Theorem 2 on the sequence $(\hat{\theta}_n)_{n\geq 1}$ is given by the next corollary.

Corollary 3 *Assume that $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (1/2, 1)$. Then, if $(\mathbf{H_S})$ and the assumptions of Proposition 2.1 hold, we have:*

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[|\hat{\theta}_n - \theta^\star|^2\right] \leq \frac{\mathrm{Tr}(\Sigma^\star)}{n} + C n^{-r_\beta}$$

where r_β is defined in Theorem 2.

2.4.2. $(L^p,\sqrt{\gamma_n})$-consistency without convexity In some many interesting cases, the latter strongly convex Assumption $\mathbf{H_{SC(\alpha)}}$ does not hold because the repelling effect towards $\theta^\star$ of $\nabla f(x)$ is not strong enough for large values of $|x|$. For example, this is the case in the logistic regression problem or in the recursive quantile estimation where the function ∇f is asymptotically flat for large values of $|x|$. Motivated by these examples, we thus aim to generalize the class of functions f for which the $(L^p,\sqrt{\gamma_n})$-consistency property holds. For this purpose, we introduce Assumption $(\mathbf{H}_\phi)$ defined by:

Assumption $(\mathbf{H}_\phi)$ - Weakly reverting drift *The function f is $\mathcal{C}^2(\mathbb{R}^d,\mathbb{R})$ with $D^2 f$ bounded and Lipschitz, $D^2 f(\theta^\star)$ invertible and:*

- *i) ϕ is $\mathcal{C}^2(\mathbb{R}_+,\mathbb{R}_+)$ non-decreasing and $\exists\, x_0 \geq 0 : \forall x \geq x_0$, $\phi''(x) \leq 0$.*
- *ii) Two positive numbers m and M exist such that $\forall x \in \mathbb{R}^d \backslash \{\theta^\star\}$:*

$$0 < m \leq \phi'(f(x))|\nabla f(x)|^2 + \frac{|\nabla f(x)|^2}{f(x)} \leq M. \tag{9}$$

Roughly speaking, the function ϕ quantifies the lack of convexity far from $\theta^\star$ and is calibrated in such a way that the function $x \longmapsto f^p(x) e^{\phi(f(x))}$ is strongly convex. The extremal situations are the following ones: when $\phi \equiv 1$, we recover the previous case or more precisely, when $x \longmapsto D^2 f(x)$ is Lipschitz continuous, $(\mathbf{H_{SC(\alpha)}}) \Longrightarrow (\mathbf{H}_\phi)$ with $\phi \equiv 1$. Actually, in this case, it is straightforward to prove that some positive constants c_1 and c_2 exist such that for all $x \in \mathbb{R}^d$,

$$\frac{c_1}{2}|x-\theta^\star|^2 \leq f(x) \leq \frac{c_2}{2}|x-\theta^\star|^2, \quad \text{and} \quad c_1|x-\theta^\star| \leq |\nabla f(x)| \leq c_2|x-\theta^\star|.$$

Note that in this case $(\mathbf{H}_\phi)$ remains slightly more general since it even can be true in some cases where $D^2 f$ is not strictly positive everywhere.

The opposite case is $\phi(x) = x$. In this setting, $(\mathbf{H}_\phi)$ is satisfied when $m \leq |\nabla f(x)|^2 \leq M$ with some positive m and M. Note that this framework includes the online logistic regression and the recursive quantile estimation (see Subsection 2.6).

For practical purposes, we introduce below a kind of parametric version of Assumption $(\mathbf{H}_\phi)$ denoted by $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$, which may be seen as a global *Kurdyka-Łojasiewicz gradient inequality* (see, *e.g.* [18, 19] and Subsection 2.5 for details):

Assumption $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ - Global KL inequality *The function f is $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ with $D^2 f$ bounded and Lipschitz, $D^2 f(\theta^\star)$ invertible and:*

- *For $r \in [0, 1/2]$, we have*

$$\liminf_{|x| \longrightarrow +\infty} f^{-r} |\nabla f| > 0 \qquad \textit{and} \qquad \limsup_{|x| \longrightarrow +\infty} f^{-r} |\nabla f| > 0 \tag{10}$$

$(\mathbf{H}_\phi)$ and $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ are linked by the following proposition:

Proposition 2.2 *Let $r \in [0, 1/2]$ such that $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ holds. Then, $(\mathbf{H}_\phi)$ holds with ϕ defined by $\phi(x) = (1 + |x|^2)^{\frac{1-2r}{2}}$. Furthermore,*

$$\liminf_{|x| \to +\infty} f(x) |x|^{-\frac{1}{1-r}} > 0. \tag{11}$$

The implication is easy to prove (using that near $\theta^\star$, $f(x) \lesssim |x - \theta^\star|^2$ and $|x - \theta^\star| \lesssim |\nabla f(x)|$ since $\nabla f(\theta^\star) = 0$ and $D^2 f(\theta^\star)$ is strictly positive). The proof of the more intricate property (11) is postponed to Appendix 6. Note that this property will be important to derive the $(L^p, \sqrt{\gamma_n})$-consistency (see Theorem 4). As mentioned before, further comments on these assumptions are postponed to Subsection 2.5 and the rest of this paragraph is devoted to the main corresponding results.

As in the strongly convex case, Assumptions $(\mathbf{H}_\phi)$ and $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ certainly need to be combined with some assumption on the martingale increment. As one might expect, the condition is unfortunately (much) more stringent than in the strongly convex case:

Assumption $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$ - Moments of the martingale increment A locally bounded deterministic function $\rho_p : \mathbb{R}_+ \mapsto \mathbb{R}_+$ exists such that:

$$\forall u \geq 0 \qquad \mathbb{E}[|\Delta M_n|^{2p+2} e^{\phi(u|\Delta M_n|^2)} | \mathcal{F}_n] \leq \rho_p(u) \qquad \text{a.s.} \tag{12}$$

Remark 2.1 *The general form of this assumption can be roughly explained as follows: one of the main ideas of the proof of Theorem 4 below is to use the function $x \mapsto f^p(x) e^{\phi(f(x))}$ as a Lyapunov-type function in order to obtain some contraction properties. Note that when $(\Delta M_n)_{n \geq 1}$ is a bounded sequence, $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$ is automatically satisfied (this is the case for the quantile recursive estimation and for the logistic regression of bounded variables: see Subsection 2.6).*

However, when $\phi \equiv 1$ (i.e. strongly convex case), it can be observed that $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$ is not retrieved as it would have been expected. This can be explained by the fact that Assumption $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$ is adapted to the general case and that the particular case $\phi \equiv 1$, certainly leads to some simplifications (especially in the derivation of the Lyapunov function). Nevertheless, we could (with additional technicalities) also allow a dependency in $f(\theta_n)$ by replacing the right-hand member of the assumption with $C(1 + (f(\theta_n))^{p-1}$. However, this seems of limited interest in the general case in view of the exponential term of the left-hand side. More precisely, the dependency in $f(\theta_n)$ could be really interesting for applications if it were of comparable size to the left-hand member. Finally, let us remark that as it can be expected, the constraint on the noise increases with ϕ, i.e., with the lack of convexity of the function f.

We then state the main result of this paragraph that holds in a generic potentially non-convex situation supported by $(\mathbf{H}_\phi)$.

Theorem 4 *For any $p \geq 1$:*
i) Assume that f satisfies $(\mathbf{H}_\phi)$ and that the martingale increment sequence satisfies $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$, then a constant C_p exists such that:

$$\mathbb{E}[f^p(\theta_n)e^{\phi(f(\theta_n))}] \leq C_p\{\gamma_n\}^p.$$

ii) If, furthermore, $\liminf_{|x|\to+\infty} |x|^{-2p} f^p(x)e^{\phi(f(x))} > 0$, then $(\theta_n)_{n\geq 1}$ is $(L^{2p},\sqrt{\gamma_n})$-consistent, e.g., a constant C_p exists such that:

$$\mathbb{E}|\theta_n - \theta^\star|^{2p} \leq C_p\{\gamma_n\}^p.$$

iii) In particular, $(\theta_n)_{n\geq 1}$ is $(L^{2p},\sqrt{\gamma_n})$-consistent if $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ holds for a given $r \in [0,1/2]$ and $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$ holds with $\phi(t) = (1+t^2)^{(1-2r)/2}$.

Proof: The proof of Theorem 4 *i)* is postponed to Section 4.

The second statement *ii)* is a simple consequence of *i)*: actually, we only need to prove that the function τ defined by $\tau(x) = f^p(x)e^{\phi(f(x))}$, $x \in \mathbb{R}^d$, satisfies $\inf_{x\in\mathbb{R}^d\setminus\{0\}} \tau(x)|x-\theta^\star|^{-2p} > 0$. Near $\theta^\star$, the fact that $D^2 f(\theta^\star)$ is positive-definite can be used to ensure that $x \mapsto \tau(x)|x-\theta^\star|^{-2p}$ is lower-bounded by a positive constant. Then, since τ is positive on $\mathbb{R}^d$, the result follows from the additional assumption $\liminf_{|x|\to+\infty} \tau(x)|x|^{-2p} > 0$.

Finally, for *iii)*, we only have to prove that the additional assumption of *ii)* holds under $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$. This point is a straightforward consequence of (11) and of the fact that $\phi(x) = (1+|x|^2)^{\frac{1-2r}{2}}$ in this case. □

Applying Theorem 2 makes it possible to derive non-asymptotic bounds under $(\mathbf{H}_\phi)$. We chose to only state the result under the parametric assumption $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$.

Corollary 5 *Assume $(\mathbf{H_S})$, $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ and $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$ with $r \in [0,1/2]$, $p = 2$ and $\phi(t) = (1+t^2)^{\frac{1-2r}{2}}$. If $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (1/2,1)$, then $(\hat{\theta}_n)_{n\geq 1}$ satisfies:*

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[|\hat{\theta}_n - \theta^\star|^2\right] \leq \frac{\mathrm{Tr}(\Sigma^\star)}{n} + Cn^{-r_\beta},$$

for C large enough and where r_β is defined in Theorem 2.

Remark 2.2 *At first sight, the result brought by Corollary (5) may appear surprising since we obtain a $O(1/n)$ rate for the mean-squared error of the averaged sequence towards $\theta^\star$* without strong convexity*, including, for example, some situations where $f(x) \sim |x|$ as $|x| \to +\infty$. This could be viewed as a contradiction with the minimax rate of convergence $O(1/\sqrt{n})$ for stochastic optimization problems in the simple convex case (see, e.g. [1] or [21]). The above minimax result simply refers to the worst situation in the class of convex functions* that are not necessarily differentiable*, whereas Assumption $(\mathbf{H}_\phi)$ used in Corollary 5 describes a set of functions that are not necessarily strongly convex or even simply convex, but all the functions involved in $(\mathbf{H}_\phi)$ or in $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ belong to $\mathcal{C}^2(\mathbb{R}^d,\mathbb{R})$ and have a positive curvature around $\theta^\star$ since we assumed that $D^2 f(\theta^\star)$ is invertible. In particular, the worst case is attained in [1] through linear combinations of shifted piecewise affine functions $x \longmapsto |x+1/2|$ and $x \longmapsto |x-1/2|$, functions for which Assumption $(\mathbf{H}_\phi)$ is obviously not satisfied. According to the results in Appendix H of [20], the local curvature near $\theta^\star$ makes it possible to obtain a $O(n^{-1})$ rate whereas the smoothness assumption allows to obtain a precise constant, leading to the Cramer-Rao lower bound in the specific setting of [20].*

2.5. Comments on Assumption ($\mathbf{H}_\phi$) and link with the Kurdyka-Łojasiewicz inequality To the best of our knowledge, this assumption is not standard in the stochastic optimization literature and thus deserves several comments, included in this section. For this purpose, for any symmetric real matrix A, let us denote the lowest eigvenvalue of A by $\underline{\lambda}_A$.

2.5.1. f does not necessarily need to be convex It is important to notice that the function f itself is not necessarily assumed to be convex under Assumption ($\mathbf{H}_\phi$). The minimal requirement is that f only possesses a unique critical point (minimum). Of course, our analysis will still be based on a descent lemma for the sequences $(\theta_n)_{n\geq 0}$. Nevertheless, we will use a Lyapunov analysis that will involve $f^p e^{\phi(f)}$ instead of f itself for the sequence $(\theta_n)_{n\geq 0}$. The descent property will then be derived from Equation (9) in *ii*) of ($\mathbf{H}_\phi$). Thereafter, we will be able to exploit a spectral analysis of the dynamical system that governs $(\hat{\theta}_n)_{n\geq 0}$. We stress the fact that, in general, the results without any convexity assumption on f are usually limited to almost sure convergence with the help of the Robbins-Siegmund Lemma (see, *e.g.* [13] and the references therein). As will be shown later on, Assumption ($\mathbf{H}_\phi$) will be sufficient to derive efficient convergence rates for the averaged sequence $(\hat{\theta}_n)_{n\geq 0}$ without any strong convexity assumption.

2.5.2. f is necessarily a sub-quadratic and L-smooth function Let us first remark that ($\mathbf{H}_\phi$) entails an a priori upper bound for f that cannot increase faster than a quadratic form. We have:

$$\forall x \in \mathbb{R}^d \qquad \frac{|\nabla f(x)|^2}{f(x)} \leq M \Longrightarrow |\nabla(\sqrt{f})| \leq \frac{\sqrt{M}}{2} \Longrightarrow f(x) \leq \frac{M}{4}\|x\|^2.$$

However, we also need a slightly stronger condition with $D^2 f$ bounded over $\mathbb{R}^d$, meaning that f is L-smooth for a suitable value of L (with an L-Lipschitz gradient). We refer to [23] for a general introduction to this class of functions. Even in the deterministic setting, the L-smooth property is a common minimal requirement for obtaining a good convergence rate for smooth optimization problems (,see, *e.g.* [4]).

2.5.3. About the Kurdyka-Łojasiewicz inequality It is important to note that ($\mathbf{H}_\phi$) should be related to the Kurdyka-Łojasiewicz gradient inequalities. In the deterministic setting, the Łojasiewicz gradient inequality [19] with exponent r may be stated as follows:

$$\exists m > 0 \quad \exists\, r \in [0,1) \quad \forall x \in \mathbb{R}^d \qquad f(x)^{-r}|\nabla f(x)| \geq m, \tag{13}$$

while a generalization (see, *e.g.* [18]) is governed by the existence of a concave increasing "desingularizing" function ψ such that:

$$|\nabla(\psi \circ f)| \geq 1.$$

The Łojasiewicz gradient inequality is then just a particular case of the previous inequality while choosing $\psi(t) = ct^{1-r}$. We refer to [6] for a recent work on how to characterize some large families of functions f such that a generalized KL-inequality holds.

In this paper, the Kurdyka-Łojasiewicz-type gradient inequality appears through Assumption ($\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}}$) with $r \in [0, 1/2]$, which implies ($\mathbf{H}_\phi$) (see Proposition 2.2). However, it should be noted that Assumption ($\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}}$) is slightly different from (13) since we only enforce the function $f^{-r}|\nabla f|$ to be *asymptotically* lower-bounded by a positive constant.

Nevertheless, in our setting where f has only one critical point and where $D^2 f(\theta^\star)$ is positive-definite, it is easy to prove that $(\mathbf{H^r_{KL}})$ implies (13). Indeed, around $\theta^\star$, $D^2 f(\theta^\star)$ is positive definite so that we could choose $r = 1/2$ and then satisfy the Łojasiewicz gradient inequality (13) near $\theta^\star$. Hence, the link between $(\mathbf{H^r_{KL}})$ given in (10) and (13) has to be understood for large values of $|x|$.

Moreover, Proposition 2.2 states that the classical Łojasiewicz gradient inequality (13) associated with the assumption of the **local** invertibility of $D^2 f(\theta^\star)$ implies Assumption $(\mathbf{H}_\phi)$. The choice $r = 1/2$ in Equation (13) corresponds to the strongly-convex case with $\phi = 1$ and $\psi(t) = \sqrt{t}$. Conversely, the Łojasiewicz exponent $r = 0$ corresponds to the weak repelling force $|\nabla f(x)|^2 \propto 1$ as $|x| \to +\infty$ and $\phi(t) = \sqrt{1+t^2}$, leading to $\psi(t) = t$.

At last, we can observe that the interest of Assumption $(\mathbf{H}_\phi)$ in the stochastic framework is more closely related to the behavior of the algorithm when $(\theta_n)_{n\geq 1}$ is far away from the target point $\theta^\star$, whereas in the deterministic framework, the main interest of the desingularizing function ψ is used around $\theta^\star$ to derive fast linear rates even in non strongly convex situations. For example, [7] established exponential convergence of the forward-backward splitting FISTA to solve the Lasso problem with the help of KL inequalities although the minimization problem is not strongly convex and the core of the study is the understanding of the algorithm near $\theta^\star$. In simple terms, the difficulty to assert some good properties of stochastic algorithms is not exactly the same as the one for deterministic problems: it is much more difficult to control the time for a stochastic algorithm to come back far away from $\theta^\star$ than for a deterministic method with a weakly reverting effect of $-\nabla f$ because of the noise on the algorithm. In contrast, the rate of a deterministic method crucially depends on the local behavior of ∇f around $\theta^\star$ (see, *e.g.* [7]).

2.5.4. Counter-examples of the global KL inequality Finally, we should have in mind what kind of functions do not satisfy the global Łojasiewicz gradient inequality given in Equation (13). Since we assumed f to have a unique minimizer $\theta^\star$ with $D^2 f(\theta^\star)$ invertible, Inequality $f^{-r}|\nabla f| \geq m > 0$ should only fail asymptotically. From Equation (11) of Proposition 2.2, we know that $|x| \lesssim f(x)$ for large values of $|x|$. As a consequence, any function f with logarithmic growth or comparable to $|x|^r$ growth with $r \in (0,1)$ at infinity can not be managed by this assumption.

Another counter-example of f occurs when f exhibits an infinite sequence of oscillations in the values of $f' \geq 0$ with longer and longer areas near $f' = 0$ when $|x|$ is increasing. We refer to [7] for the following function that does not satisfy the KL inequality for any $r \geq 2$:

$$f : x \longrightarrow x^{2r}[2 + \cos(x^{-1})] \quad \text{if} \quad x \neq 0 \quad \text{and} \quad f(0) = 0.$$

2.6. Applications

2.6.1. Strongly convex situation First, we can observe that in the strongly convex situation, Corollary 3 provides a very tractable criterion to assess the non-asymptotic first-order optimality of the averaging procedure since $(\mathbf{H^{SC}_{\Sigma_p}})$ is very easy to check.

For example, considering the **stochastic recursive least mean square estimation** problem (see, *i.e.*, [13]), it can immediately be checked that $\theta \longrightarrow f(\theta)$ is quadratic. In that case, the problem is strongly convex, and the noise increment satisfies:

$$\mathbb{E}[|\Delta M_n|^{2p}|\mathcal{F}_n] \leq \Sigma_p(1 + (f(\theta_n))^p \qquad \text{a.s.}$$

Then Proposition 2.1 yields the $(L^p, \sqrt{\gamma_n})$ consistency rate of $(\theta_n)_{n\geq 1}$, which implies a first-order optimal excess risk for $(\hat{\theta}_n)_{n\geq 1}$ with a $O(n^{-5/4})$ second-order term. We stress the fact that the recent contribution of [3] also proves a sharp non-asymptotic $O(1/n)$ rate of convergence with a $O(n^{-7/6})$ second-order term. Hence, Corollary 3 yields a stronger result in that case.

2.6.2. Assumptions $(\mathbf{H}_\phi)$ and $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$ hold for many stochastic minimization problems
We end this section by pointing out that Assumption $(\mathbf{H}_\phi)$ and $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$ capture many interesting situations where f is not strongly convex and may even not be convex in some cases.

2.6.2.a. Semi-algebraic case

Before providing explicit examples, a general argument relies on the statement of Theorem 2 of [6]: every coercive convex continuous function f, which is proper and semi-algebraic (see [6] for some precise definitions), satisfies the KL inequality. Note that such a result holds in non-smooth situations, as stated in [5], when using sub-differential instead of gradients, but our work does not deal with non smooth-functions f.

2.6.2.b. On-line logistic regression

The on-line logistic regression problem deals with the minimization of f defined by:

$$f(\theta) := \mathbb{E}\left[\log\left(1+e^{-Y<X,\theta>}\right)\right] \tag{14}$$

where X is a $\mathbb{R}^d$ random variable and $Y|X$ takes its value in $\{-1,1\}$ with:

$$P[Y=1\,|X=x] = \frac{1}{1+e^{-<x,\theta^\star>}}. \tag{15}$$

We then observe a sequence of i.i.d. replications (X_i,Y_i) and the baseline stochastic gradient descent sequence $(\theta_n)_{n\geq 1}$ is defined by:

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\frac{Y_n X_n}{1+e^{Y_n<\theta_n,X_n>}} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta M_{n+1}. \tag{16}$$

We state the following result below.

Proposition 2.3 *Assume that the law of the design X is compactly supported in $B_{\mathbb{R}^d}(0,R)$ for a given $R>0$ and is elliptic: for any $e\in\mathcal{S}^{d-1}(\mathbb{R}^d)$, $Var(<X,e>)\geq 0$. Assume that Y satisfies the logistic Equation (15). Then*

i) f defined in Equation (14) is convex with $D^2 f$ bounded and Lipschitz. Moreover $D^2 f(\theta^\star)$ is invertible and satisfies $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ with $r=0$.

ii) Recall that $\Sigma^\star$ is defined in (6), the averaged sequence $(\hat{\theta}_n)_{n\geq 1}$ built from the sequence $(\theta_n)_{n\geq 1}$ introduced in (16) satisfies:

$$\exists C>0 \quad \forall n\geq 1 \qquad \mathbb{E}|\hat{\theta}_n-\theta^\star|^2 \leq \frac{Tr(\Sigma^\star)}{n} + Cn^{-5/4}.$$

Proof: We study i). Some straightforward computations yield $\forall\theta\in\mathbb{R}^d$:

$$\nabla f(\theta) = \mathbb{E}\left[\frac{X\left[e^{<X,\theta>}-e^{<X,\theta^\star>}\right]}{\left[1+e^{<X,\theta>}\right]\left[1+e^{<X,\theta^\star>}\right]}\right] \quad \text{and} \quad D^2 f(\theta)_{k,l} = \mathbb{E}\left[\frac{X_k X_l e^{<X,\theta>}}{(1+e^{<X,\theta>})^2}\right]$$

We can deduce that $\nabla f(\theta^\star)=0$ and that (see [2] for example) f is convex with

$$<\theta-\theta^\star, \nabla f(\theta)> = \mathbb{E}\left[\frac{\left[<X,\theta>-<X,\theta^\star>\right]\left[e^{<X,\theta>}-e^{<X,\theta^\star>}\right]}{\left[1+e^{<X,\theta^\star>}\right]\left[1+e^{<X,\theta>}\right]}\right] \geq 0,$$

because $(x-y)[e^x-e^y]>0$ for every pair (x,y) such that $x\neq y$. It implies that $\theta^\star$ is the unique minimizer of f. Moreover, $D^2f(\theta^\star)=\mathbb{E}\left[XX^T\frac{e^{<X,\theta^\star>}}{(1+e^{<X,\theta^\star>}}\right]$ is invertible as soon as the design matrix is invertible. This property easily follows from the ellipticity condition on the distribution of the design:

$$\forall e\in\mathcal{S}^{d-1}(\mathbb{R}^d)\qquad Var(<X,e>)=e^T\mathbb{E}[XX^T]e>0,$$

which proves that the Hessian $D^2f(\theta^\star)$ is invertible.

Regarding now the asymptotic norm of $|\nabla f(\theta)|$, the Lebesgue Theorem yields, $\forall e\in\mathcal{S}^{d-1}(\mathbb{R}^d)$:

$$\begin{aligned}\lim_{t\longrightarrow+\infty}|\nabla f(te)| &= \left|\mathbb{E}\left[\frac{X\mathbf{1}_{<X,e>\geq 0}-Xe^{<X,\theta^\star>}\mathbf{1}_{<X,e><0}}{1+e^{<X,\theta^\star>}}\right]\right|\\ &= \left|\left\langle\mathbb{E}\left[\frac{X\mathbf{1}_{<X,e>\geq 0}-Xe^{<X,\theta^\star>}\mathbf{1}_{<X,e><0}}{1+e^{<X,\theta^\star>}}\right],e\right\rangle\right|\\ &\geq \left|\mathbb{E}\left[\frac{<X,e>\mathbf{1}_{<X,e>\geq 0}-<X,e>e^{<X,\theta^\star>}\mathbf{1}_{<X,e><0}}{1+e^{<X,\theta^\star>}}\right]\right|\\ &\geq \left|\mathbb{E}\left[\frac{<X,e>\mathbf{1}_{<X,e>\geq 0}}{1+e^{<X,\theta^\star>}}\right]\right|\wedge\left|\mathbb{E}\left[\frac{<X,-e>e^{<X,\theta^\star>}\mathbf{1}_{<X,-e>\geq 0}}{1+e^{<X,\theta^\star>}}\right]\right|\end{aligned}$$

where we used the orthogonal decomposition on e and $e^\perp$. It then proves that for any $e\in\mathcal{S}^{d-1}(\mathbb{R}^d)$, $\lim_{t\longrightarrow+\infty}|\nabla f(te)|>0$. A compactness and continuity argument leads to:

$$\liminf_{|\theta|\longrightarrow+\infty}|\nabla f(\theta)|\geq\frac{\inf_{e\in\mathcal{S}^{d-1}(\mathbb{R}^d)}\mathbb{E}\left[<X,e>_+\right]}{e^{R|\theta^\star|}(1+e^{R|\theta^\star|})}>0,$$

since we assumed the design to be elliptic: $Var(<X,e>)>0$ for any unit vector e. At the same time, it is also straightforward to check that:

$$\limsup_{|\theta|\longrightarrow+\infty}|\nabla f(\theta)|\leq+\infty,$$

which concludes the proof of i).

We now prove ii) and apply Corollary 5. In that case, Assumption $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ holds with $r=0$. Regarding Assumption $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$, we can observe that the martingale increments are *bounded* (see [2], for example) and Inequality (12) is satisfied. Hence, Corollary 5 implies that $(\theta_n)_{n\geq1}$ is a L^p-$\{\sqrt{\gamma_n}\}$ consistent sequence for any $p\geq2$. We can therefore apply Theorem 2 for the averaging procedure $(\hat{\theta}_n)_{n\geq1}$, with $\Sigma^\star$ given in (6). This ends the proof. □

2.6.2.c. Recursive quantile estimation

The recursive quantile estimation problem is a standard example that may be stated as follows (see, *e.g.* [13] for details). For a given cumulative distribution function G defined over $\mathbb{R}$, the problem is to find the quantile q_α such that $G(q_\alpha)=1-\alpha$. We assume that we observe a sequence of i.i.d. realizations $(X_i)_{i\geq1}$ distributed with a cumulative distribution G. The recursive quantile algorithm is then defined by:

$$\theta_{n+1}=\theta_n-\gamma_{n+1}\left[\mathbf{1}_{X_n\leq\theta_n}-(1-\alpha)\right]=\theta_n-\gamma_{n+1}[G(\theta_n)-(1-\alpha)]+\gamma_{n+1}\Delta M_{n+1},$$

In that situation, the function f' is defined by:

$$f'(\theta)=\int_{q_\alpha}^{\theta}p(s)ds=G(\theta)-G(q_\alpha),$$

where p is the density with respect to the Lebesgue measure such that $G(q) = \int_{-\infty}^{q} p$. Below, we consider the case where p is a Lipschitz continuous function with $p(q_\alpha) > 0$. Assuming without loss of generality that $q_\alpha = 0$ so that $f(0) = 0$, the function f is then defined by:

$$f(\theta) := \int_0^\theta \int_0^u p(s) ds du,$$

whose minimum is attained at 0. It can immediately be checked that $f''(0) \neq 0$ as soon as $p(q_\alpha) > 0$ and $f'(\theta) \longrightarrow 1-\alpha$ when $\theta \longrightarrow +\infty$ while $f'(\theta) \longrightarrow -\alpha$ when $\theta \longrightarrow -\infty$. Therefore, f satisfies $(\mathbf{H}_\phi)$ since $(\mathbf{H}^r_{KL})$ and Equation (13) hold with $r = 0$ and $\phi(t) = \sqrt{1+t^2}$. Again, regarding Assumption $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$, we can observe that the martingale increments are *bounded* (see [10, 13], for example). Therefore, Inequality (12) is obviously satisfied since ϕ is a monotone increasing function. We can apply Corollary 5 and conclude that the averaging sequence $(\hat{\theta}_n)_{n\geq 1}$ satisfies the non-asymptotic optimal inequality: a constant $C > 0$ exists such that:

$$\forall n \geq 1 \qquad \mathbb{E}|\hat{\theta}_n - q_\alpha|^2 \leq \frac{\alpha(1-\alpha)}{p(q_\alpha)\, n} + C n^{-5/4}$$

• The **on-line geometric median estimation** We end this section with considerations on a problem close to the former one in larger dimensional spaces. The median estimation problem described in [10, 9] relies on the minimization of:

$$\forall \theta \in \mathbb{R}^d \qquad f(\theta) = \mathbb{E}[|X - \theta|],$$

where X is a random variable distributed over $\mathbb{R}^d$. Of course, our framework does not apply to this situation since f is not $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$. Nevertheless, if we assume for the sake of simplicity that the support of X is bounded (which is not assumed in the initial works of [10, 9]), then following the arguments of [17], the median is uniquely defined as soon as the distribution of X is not concentrated on a single straight line, meaning that the variance of X is elliptic in any direction of the sphere of $\mathbb{R}^d$. Moreover, it can be easily seen that:

$$\lim_{|\theta| \longrightarrow +\infty} |\nabla f(\theta)| = 1,$$

so that Equation (13) holds with $r = 0$. To apply Corollary 5, it would be necessary to extend our work to this *non-smooth* situation, which is beyond the scope of this paper, but that would be an interesting future subject of investigation.

2.7. Organization of the paper The rest of the paper is dedicated to the proofs of the main results and the text is then organized as follows. We first assume without loss of generality that $\theta^\star = 0$ (and that $f(\theta^\star) = 0$). In Section 3, we detail our spectral analysis of the behavior of $(\hat{\theta}_n)_{n\geq 1}$ and prove Theorem 2. In particular, Proposition 3.4 provides the main argument to derive the sharp exact first-order rate of convergence, and the results postponed below in Section 3 only represent technical lemmas that are useful for the proof of Proposition 3.4. Section 4 is dedicated to the proof of the $(L^p, \sqrt{\gamma_n})$-consistency under Assumption $(\mathbf{H}_\phi)$ (proof of Theorem 4 *i*)). The generalization to the stronger situation of strong convexity (Proposition 2.1) is left to the reader since it only requires slight modifications of the proof).

3. Non asymptotic optimal averaging procedure

3.1. Proof of Theorem 2 The aim of this paragraph is to prove Theorem 2.

We will use a coupled relationship between $\hat{\theta}_{n+1}$ and $(\hat{\theta}_n, \theta_{n+1})$. For this purpose, we introduce the notation for the drift at time n:

$$\Lambda_n := \int_0^1 D^2 f(t\theta_n) dt \quad \text{so that} \quad \Lambda_n \theta_n = \nabla f(\theta_n) \tag{17}$$

using the Taylor formula and the fact that $\theta^\star = \nabla f(\theta^\star) = 0$. The coupled evolution $(\theta_n, \hat{\theta}_n) \to (\theta_{n+1}, \hat{\theta}_{n+1})$ is then described by the next proposition.

Proposition 3.1 *If we now introduce $Z_n = (\theta_n, \hat{\theta}_n)$, then we have the $2d$-dimensional recursion formula:*

$$Z_{n+1} = \begin{pmatrix} I_d - \gamma_{n+1}\Lambda_n & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1}\Lambda_n) & (1 - \frac{1}{n+1})I_d \end{pmatrix} Z_n + \gamma_{n+1} \begin{pmatrix} \Delta M_{n+1} \\ \frac{\Delta M_{n+1}}{n+1} \end{pmatrix}. \tag{18}$$

Proof: We begin with the simple remark:

$$\forall n \in \mathbb{N} \qquad \hat{\theta}_{n+1} = \hat{\theta}_n + \frac{1}{n+1}\left(\theta_{n+1} - \hat{\theta}_n\right).$$

Now, Equation (2) yields:

$$\forall n \in \mathbb{N} \qquad \begin{cases} \theta_{n+1} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta M_{n+1} \\ \hat{\theta}_{n+1} = \hat{\theta}_n(1 - \frac{1}{n+1}) + \frac{1}{n+1}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta M_{n+1}\right). \end{cases}$$

The result then follows from (17). □

The next proposition describes the linearization procedure by replacing Λ_n with the fixed Hessian of f at $\theta^\star$.

Proposition 3.2 *Set $\Lambda^\star = D^2 f(\theta^\star)$ and assume that $\Lambda^\star$ is a positive-definite matrix. Then, a matrix $Q \in \mathcal{O}_d(\mathbb{R})$ exists such that $\check{Z}_n = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} Z_n$ satisfies:*

$$\check{Z}_{n+1} = A_n \check{Z}_n + \gamma_{n+1} \begin{pmatrix} Q\Delta M_{n+1} \\ \frac{Q\Delta M_{n+1}}{n+1} \end{pmatrix} + \underbrace{\gamma_{n+1} \begin{pmatrix} Q(\Lambda^\star - \Lambda_n)\theta_n \\ Q(\Lambda_n - \Lambda^\star)\frac{\theta_n}{n+1} \end{pmatrix}}_{:=\check{\upsilon}_n}, \tag{19}$$

where $D^\star$ is the diagonal matrix associated with the eigenvalues of $\Lambda^\star$ and

$$A_n := \begin{pmatrix} I_d - \gamma_{n+1}D^\star & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1}D^\star) & (1 - \frac{1}{n+1})I_d \end{pmatrix}. \tag{20}$$

Proof: We write $\Lambda_n = \underbrace{D^2 f(\theta^\star)}_{:=\Lambda^\star} + (\Lambda_n - D^2 f(\theta^\star))$ and use the eigenvalue decomposition of $\Lambda^\star$.

$$Z_{n+1} = \begin{pmatrix} I_d - \gamma_{n+1}\Lambda^\star & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1}\Lambda^\star) & (1 - \frac{1}{n+1})I_d \end{pmatrix} Z_n + \gamma_{n+1} \begin{pmatrix} \Delta M_{n+1} \\ \frac{\Delta M_{n+1}}{n+1} \end{pmatrix} + \upsilon_n, \tag{21}$$

where the linearization term υ_n will be shown to be negligible and is defined by

$$\upsilon_n := \gamma_{n+1} \begin{pmatrix} (\Lambda^\star - \Lambda_n)\theta_n \\ (\Lambda_n - \Lambda^\star)\frac{\theta_n}{n+1} \end{pmatrix}.$$

The matrix $\Lambda^\star$ is the Hessian of f at $\theta^\star$ and is a symmetric positive matrix, which may be reduced into a diagonal matrix $D^\star = Diag(\mu_1^\star, \ldots, \mu_d^\star)$ with positive eigenvalues in an orthonormal basis:

$$\exists Q \in \mathcal{O}_d(\mathbb{R}) \qquad \Lambda^\star = Q^T D^\star Q \qquad \text{with} \qquad Q^T = Q^{-1}. \tag{22}$$

It is natural to introduce the new sequence adapted to the spectral decomposition of $\Lambda^\star$ given by Equation (22):

$$\check{Z}_n = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} Z_n = \begin{pmatrix} Q\theta_n \\ Q\hat{\theta}_n \end{pmatrix}. \tag{23}$$

Using $Q\Lambda^\star = D^\star Q$, we obtain the equality described in Equation (19). □ The important fact about the evolution of $(\check{Z}_n)_{n\geq 1})$ is the blockwise structure of A_n as d blocks of 2×2 matrices:

$$A_n = \begin{pmatrix} \begin{bmatrix} 1-\gamma_{n+1}\mu_1^\star & 0 & \ldots & 0 \\ 0 & 1-\gamma_{n+1}\mu_2^\star & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1-\gamma_{n+1}\mu_d^\star \end{bmatrix} & \mathbf{0_d} \\ \begin{bmatrix} \frac{1-\gamma_{n+1}\mu_1^\star}{n+1} & 0 & \ldots & 0 \\ 0 & \frac{1-\gamma_{n+1}\mu_2^\star}{n+1} & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & \frac{1-\gamma_{n+1}\mu_d^\star}{n+1} \end{bmatrix} & (1-\frac{1}{n+1})\mathbf{I_d} \end{pmatrix}. \tag{24}$$

In particular, we can observe that the matrices made of components (i,i) $(i,d+i)$, $(d+i,i)$ and $(d+i,d+i)$ have a similar form. In the next proposition, we focus on the related spectrum of such 2×2-matrices (the proof is left to the reader).

Proposition 3.3 *For $\mu \in \mathbb{R}$ and $n \geq 1$, set $E_{\mu,n} := \begin{pmatrix} 1-\gamma_{n+1}\mu & 0 \\ \frac{1-\mu\gamma_{n+1}}{n+1} & 1-\frac{1}{n+1} \end{pmatrix}$.* • *If $1-\mu\gamma_{n+1}(n+1) \neq 0$, define $\epsilon_{\mu,n+1}$ by:*

$$\epsilon_{\mu,n+1} := \frac{1-\mu\gamma_{n+1}}{1-\mu\gamma_{n+1}(n+1)}, \tag{25}$$

The eigenvalues of $E_{\mu,n}$ are then given by

$$Sp(E_{\mu,n}) = \left\{1-\mu\gamma_{n+1}, 1-\frac{1}{n+1}\right\},$$

whereas the associated eigenvectors are:

$$u_{\mu,n} = \begin{pmatrix} 1 \\ \epsilon_{\mu,n+1} \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

• *If $1-\mu\gamma_{n+1}(n+1) = 0$, $E_{\mu,n}$ is not diagonalizable in $\mathbb{R}$.*

At this stage, we point out that the eigenvectors are modified from one iteration to another in our spectral analysis of $(\hat{\theta}_n)_{n\geq 1}$. Lemma 5 (stated in Appendix 5 will be useful to assert how much the eigenvectors are moving.

Remark 3.1 *The spectral decomposition of $E_{\mu,n}$ will be important below.*

• *The first important remark is that $E_{\mu,n}$ is not symmetric. The same remark holds for A_n as well as shown in Equation (24). This generates a non-orthonormal change of basis to reduce $E_{\mu,n}$ and A_n into a diagonal form, which implies some technical complications for the study of $(\check{Z}_n)_{n\geq 1}$.*

- *To a lesser extent, it is also interesting to point out that this "no self-adjointness" property of A_n is a new example of acceleration of convergence rates with the help of non symmetric dynamical systems. This phenomenon also occurs for the kinetic diffusion dynamics [27, 15]) and for the Nesterov accelerated gradient descent [22] and the Heavy Ball system [8, 16] even though we do not claim that such a clear common point exists between these methods.*
- *The first eigenvalue of $E_{\mu,n}$ is $1-\mu\gamma_{n+1}$, and essentially acts on the component θ_n of the vector Z_n. We then expect a contraction of θ_n related to $\prod_{k=1}^{n}(1-\mu\gamma_{k+1})$ where μ is the associated eigenvalue of the Hessian of f at $\theta^\star$. In a sense, there is nothing new for the standard stochastic gradient descent algorithm in this last observation.*
- *Interestingly, the second eigenvalue of $E_{\mu,n}$ is $1-(n+1)^{-1}$, which is* **independent** *of the value of μ. Moreover, this eigenvalue acts on the component brought by $\hat{\theta}_n$ in the vector Z_n. This key observation will be at the core of the argument for a non-asymptotic study of the Ruppert-Polyak algorithm and an important fact to obtain the adaptivity property for the unknown value of $D^\star$. In the following section, we obtain some helpful properties on the averaging procedure due to a careful inspection of the evolution of the eigenvalues of $E_{\mu,n}$ from n to $n+1$.*

The reduction of $E_{\mu,n}$ may be written as:

$$E_{\mu,n} = \begin{pmatrix} 1 & 0 \\ \epsilon_{\mu,n+1} & 1 \end{pmatrix} \begin{pmatrix} 1-\mu\gamma_{n+1} & 0 \\ 0 & 1-\frac{1}{n+1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\epsilon_{\mu,n+1} & 1 \end{pmatrix}.$$

Therefore, if we define the diagonal matrix $\mathcal{E}_{n,D^\star}$ by:

$$\mathcal{E}_{n,D^\star} = Diag(\epsilon_{\mu_1^\star,n+1},\ldots,\epsilon_{\mu_d^\star,n+1}), \tag{26}$$

we then deduce the spectral decomposition of A_n:

$$A_n = \begin{pmatrix} I_d & 0 \\ \mathcal{E}_{n,D^\star} & I_d \end{pmatrix} \begin{pmatrix} I_d-\gamma_{n+1}D^\star & 0 \\ 0 & (1-\frac{1}{n+1})I_d \end{pmatrix} \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n,D^\star} & I_d \end{pmatrix}. \tag{27}$$

We introduce the last change of basis as:

$$\widetilde{Z}_n := \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n,D^\star} & I_d \end{pmatrix} \check{Z}_n. \tag{28}$$

We will establish the following proposition.

Proposition 3.4 *Assume that $\Lambda^\star$ is a positive-definite matrix. If $(\theta_n)_{n\geq 1}$ is a $(L^p,\sqrt{\gamma_n})$-consistent sequence with $p\geq 4$ and if $(\mathbf{H_S})$ holds then the sequence $(\widetilde{Z}_n)_{n\geq 0}=(\widetilde{Z}_n^{(1)},\widetilde{Z}_n^{(2)})_{n\geq 0}$ satisfies:*

- *i) Some constants $(c_p)_{p\geq 1}$ exists such that:*

$$\forall n\geq 1 \qquad \mathbb{E}\left|\widetilde{Z}_n^{(1)}\right|^p \lesssim c_p\{\gamma_n\}^{\frac{p}{2}}.$$

- *ii) A constant c_2 exists such that:*

$$\forall n\geq 1 \qquad \mathbb{E}\left|\widetilde{Z}_n^{(2)}\right|^2 \leq \frac{\mathrm{Tr}(\Sigma^\star)}{n}+\frac{c_2}{n^{r_\beta}},$$

where $r_\beta=\{(\beta+1/2)\wedge(2-\beta)\}>1$ as soon as $\beta\in(1/2,1)$.

Since we aim to obtain the highest possible value for the second order term r_β, we are driven to the "optimal" choice $\beta = 3/4$, which in turns implies that

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}|\widetilde{Z}_n|_2^2 \leq \frac{\mathrm{Tr}(\Sigma^\star)}{n} + Cn^{-5/4}.$$

Proof:

Proof of *i*): We first observe that the sequence in $\mathbb{R}^d \times \mathbb{R}^d$ may be written as $\widetilde{Z}_n = (\widetilde{Z}_n^{(1)}, \widetilde{Z}_n^{(2)})$ and Equations (23) and (28) prove that $\widetilde{Z}_n^{(1)} = Q\theta_n$. Then, the $(L^p, \sqrt{\gamma_n})$-consistency of $(\widetilde{Z}_n^{(1)})_{n\geq 1}$ is a direct consequence of the one of $(\theta_n)_{n\geq 1}$.

Proof of *ii*): We pick n_0 such that $\forall n \geq n_0 : \epsilon_{\mu,n} < 0$ for any $\mu \in Sp(\Lambda^\star)$.

Step 1: Recursion formula

We first establish a recursion between $\widetilde{Z}_n$ and $\widetilde{Z}_{n+1}$ that will be used in Lemma 6. It will provide a key relationship on the covariance between $\widetilde{Z}_n^{(1)}$ and $\widetilde{Z}_n^{(2)}$ and on the variance of $\widetilde{Z}_n^{(2)}$.

Definitions (23), (28), the recursive link (21) and the definition of $\check{\upsilon}_n$ given in Equation (19) yield:

$$\begin{aligned}
\widetilde{Z}_{n+1} &= \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n+1,D^\star} & I_d \end{pmatrix} \check{Z}_{n+1} \\
&= \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n+1,D^\star} & I_d \end{pmatrix} \left(A_n \check{Z}_n + \gamma_{n+1} \begin{pmatrix} Q\Delta M_{n+1} \\ \frac{Q\Delta M_{n+1}}{n+1} \end{pmatrix} + \check{\upsilon}_n \right) \\
&= \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n+1,D^\star} & I_d \end{pmatrix} \begin{pmatrix} I_d & 0 \\ \mathcal{E}_{n,D^\star} & I_d \end{pmatrix} \begin{pmatrix} I_d - \gamma_{n+1} D^\star & 0 \\ 0 & (1 - \frac{1}{n+1}) I_d \end{pmatrix} \widetilde{Z}_n \\
&+ \gamma_{n+1} \left[\begin{pmatrix} Q\Delta M_{n+1} \\ (-\mathcal{E}_{n+1,D^\star} + \frac{I_d}{n+1}) Q\Delta M_{n+1} \end{pmatrix} + \begin{pmatrix} Q(\Lambda^\star - \Lambda_n)\theta_n \\ (\mathcal{E}_{n+1,D^\star} - \frac{I_d}{n+1}) Q(\Lambda^\star - \Lambda_n)\theta_n \end{pmatrix} \right],
\end{aligned}$$

where in the third line we used the spectral decomposition of A_n given by (27). Since $D^2 f$ is Lipschitz continuous, $\|\Lambda^\star - \Lambda_n\| = O(|\theta_n|)$. Then, we deduce that:

$$\begin{cases} \widetilde{Z}_{n+1}^{(1)} = (I_d - \gamma_{n+1} D^\star) \widetilde{Z}_n^{(1)} + \gamma_{n+1} \left(Q\Delta M_{n+1} + O\left(|\theta_n|^2\right) \right) \\ \widetilde{Z}_{n+1}^{(2)} = (1 - \frac{1}{n+1}) \widetilde{Z}_n^{(2)} + \Omega_n \widetilde{Z}_n^{(1)} + \gamma_{n+1} \Upsilon_n \left(Q\Delta M_{n+1} + O\left(|\theta_n|^2\right) \right), \end{cases} \tag{29}$$

with

$$\Omega_n = (\mathcal{E}_{n,D^\star} - \mathcal{E}_{n+1,D^\star})(I_d - \gamma_{n+1} D^\star) \quad \text{and} \quad \Upsilon_n = \mathcal{E}_{n+1,D^\star} - \frac{I_d}{n+1}.$$

Step 2: $\mathbb{E}[|\tilde{\mathbf{Z}}_{\mathbf{n}}^{(\mathbf{2})}|^{\mathbf{2}}] = \mathbf{O}(\mathbf{n}^{-\mathbf{1}})$ The study of $\mathbb{E}[|\theta_n|^2 \tilde{Z}_n^{(2)}]$ is rather intricated as pointed in Lemma 6. We introduce the covariance:

$$\forall i \in \{1, \ldots, d\} \qquad \omega_n(i) = \mathbb{E}[(\widetilde{Z}_n)_i (\widetilde{Z}_n)_{d+i}] = \mathbb{E}[(\widetilde{Z}_n^{(1)})_i (\widetilde{Z}_n^{(2)})_i], \tag{30}$$

and the useful coefficient:

$$\forall i \in \{1, \ldots, d\} \qquad \alpha_n^i = 2 \left(1 - \frac{1}{n+1} \right) \{\Omega_n\}_{i,i}. \tag{31}$$

We can use the Young inequality $ab \leq \frac{\epsilon}{2} a^2 + \frac{1}{2\epsilon} b^2$ with some well-chosen ϵ. More precisely, setting $\epsilon = n^r$, we obtain:

$$\mathbb{E}[|\theta_n|^2 |\tilde{Z}_n^{(2)}|] \lesssim n^r \mathbb{E}[|\theta_n|^4] + n^{-r} \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] \leq n^{r-2\beta} + n^{-r} \mathbb{E}[|\tilde{Z}_n^{(2)}|^2].$$

Since $2\beta > 1$, we know that a $\delta > 0$ exists such that $r = 2\beta - 1 - \delta > 0$ and

$$\frac{\mathbb{E}[|\theta_n|^2 |\tilde{Z}_n^{(2)}|]}{n} \leq n^{-2-\delta} + n^{-2\beta+\delta} \mathbb{E}|\tilde{Z}_n^{(2)}|^2.$$

Second, from Lemma 6, for every $i \in \{1,\dots,d\} : |\alpha_n^i| \lesssim \{n^2\gamma_n\}^{-1}$ and

$$|\alpha_n^i \omega_n(i)| \lesssim \frac{1}{\gamma_n n^2}\left(n^{-\delta'}\mathbb{E}|\widetilde{Z}_n^{(1)}|^2 + n^{\delta'}\mathbb{E}|\tilde{Z}_n^{(2)}|^2\right)$$
$$\leq n^{-2-\delta'} + n^{\beta+\delta'-2}\mathbb{E}|\tilde{Z}_n^{(2)}|^2.$$

Plugging the two previous controls into the second statement of Lemma 6, we get a positive δ such that a n_0 exists such that for all $n \geq n_0$:

$$\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] \leq \left(\left(1-\frac{1}{n+1}\right)^2 + C[n^{-2\beta+\delta} + n^{\beta+\delta'-2}]\right)\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + \frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2}$$
$$+C\left(n^{-(2+\delta)} + n^{-(2+\delta')} + n^{-(2+\beta/2)} + n^{-3+\beta}\right).$$

We choose $\delta = \beta - 1/2 > 0$ and $\delta' = 1/2 - \beta/2 > 0$. In the meantime, we also have $2+\delta\wedge 2+\delta'\wedge 2+\beta/2\wedge 3-\beta > 2$. According to this choice, we can apply Lemma 8 and deduce that a $\eta > 0$ exists such that:

$$\forall n \geq 1 \qquad \mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] \leq \frac{\mathrm{Tr}(\Sigma^\star)}{n+1}\left(1+O(n^{-\eta})\right).$$

Step 3: Control of the covariance Owing to the previous control of $\mathbb{E}[|\tilde{Z}_n^{(2)}|^2]$, one can deduce from Cauchy-Schwarz inequality that:

$$\mathbb{E}[|\theta_n|^2|\tilde{Z}_n^{(2)}|] \leq \sqrt{\mathbb{E}[|\theta_n|^4]}\sqrt{\mathbb{E}[|\tilde{Z}_n^{(2)}|^2]} \lesssim \frac{\gamma_n}{\sqrt{n}}. \tag{32}$$

Plugging this control into Lemma 6 *i*), we obtain that for all $i \in \{1,\dots,d\}$:

$$\omega_{n+1}(i) = (1-\gamma_{n+1}\mu_i^\star)\frac{n}{n+1}\omega_n(i) + O\left(\frac{\gamma_{n+1}}{n+1}\right) + O\left(\frac{\gamma_{n+1}^2}{\sqrt{n}}\right).$$

Now, remark that $\gamma_n \lesssim \sqrt{n}$ so that we can conclude that $\mathbb{E}[|\theta_n|^2|\tilde{Z}_n^{(2)}|]$ shall be neglected in the evolution of $(\omega_n(i))_{n\geq 1}$:

$$\omega_{n+1}(i) = (1-\gamma_{n+1}\mu_i^\star)\frac{n}{n+1}\omega_n(i) + O\left(\frac{\gamma_{n+1}}{n+1}\right).$$

From Lemma 7 stated in Appendix 5, we conclude that:

$$\forall i \in \{1,\dots,d\} \qquad \omega_n(i) = O\left(\frac{1}{n}\right). \tag{33}$$

Step 4: Expansion of the quadratic error We can conclude the proof of Proposition 3.4 *ii*). From the previous upper bounds (33) and (32), we have:

$$\sum_{i=1}^d \alpha_n^i \omega_n(i) = O\left(\frac{1}{n^2\gamma_n}\right) \times O\left(\frac{1}{n}\right) \quad \text{and} \quad \frac{\mathbb{E}[|\theta_n|^2|\tilde{Z}_n^{(2)}|]}{n} = O\left(\frac{\gamma_n}{n\sqrt{n}}\right).$$

We use these bounds in the statement of Lemma 6 *ii*) and deduce that:

$$\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] \leq \left(1-\frac{1}{n+1}\right)^2\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + O\left(\frac{1}{n^3\gamma_n}\right) + O\left(\frac{\gamma_n}{n^{\frac{3}{2}}}\right) + O\left(\frac{\sqrt{\gamma_n}}{n^2}\right)$$
$$\leq \left(1-\frac{1}{n+1}\right)^2\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + O\left(\frac{1}{n^{(\frac{3}{2}+\beta)\wedge(3-\beta)}}\right)$$

where we used that $\gamma_n = \gamma n^{-\beta}$ so that $\sqrt{\gamma_n}n^{-2} = o(\gamma_n n^{-3/2})$ regardless the value of $\beta \in (1/2,1)$. Applying again Lemma 8 with $r = +\infty$ and $q_\beta = (\frac{3}{2}+\beta)\wedge(3-\beta)$, one obtains the announced result. □

3.2. Further remarks on the second order term

Remark 3.2 (About the linear case) *When $x \mapsto D^2 f(x)$ is constant (or also when the function f to minimize is $\mathcal{C}^3$ with third partial derivatives Lipschitz and null at $\theta^\star$), we can remark that $\Lambda_n = \Lambda^\star$ (or that $\Lambda_n - \Lambda^\star = O(|\theta_n|^2)$). Following carefully the proof of Lemma 6, we can deduce that the error term $n^{-1}O(\mathbb{E}[|\theta_n|^2|\tilde{Z}_n^{(2)}|])$ vanishes (or is replaced by $n^{-1}O(\mathbb{E}[|\theta_n|^3|\tilde{Z}_n^{(2)}|]) \lesssim (n^{-1}\gamma_n)^{\frac{3}{2}}$ if the $(L^6, \sqrt{\gamma_n})$-consistency holds). Hence, we obtain*

$$\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] \leq \left(1 - \frac{1}{n+1}\right)^2 \mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + O(n^{-3}\gamma_n^{-1}) + O\left(\frac{\sqrt{\gamma_n}}{n^2}\right),$$

which is a better upper bound comparing to the recursion obtained in the end of the previous proof. The rate is then optimized by choosing $\beta = 2/3$, leading to an exponent $n^{-\frac{4}{3}}$.

The previous remark shows that we may obtain a different size of the second order terms when f is locally symmetric around $\theta^\star$ (which occurs when $D^3 f(\theta^\star) = 0$) and when f is not locally symmetric (Theorem 2 proves that this second order term may be fixed of size $O(n^{-5/4})$). To confirm such a conjecture, we have computed with a Monte-Carlo approximation the evolution of $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n}\right)$ with $\rho = \frac{5}{4}$ and $\beta = \frac{3}{4}$ for a locally non-symmetric f_1 around $\theta^\star$ and $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n}\right)$ with $\rho = \frac{4}{3}$ and $\beta = \frac{2}{3}$ for a locally symmetric f_2.

We have used $f_1(x) = \frac{x^2}{2} e^{-\arctan(x)}$ and $f_2(x) = \frac{x^2}{2}$, which trivially fall in the two different cases and the simulations illustrated by Figure 1 seem to confirm that the second-order terms are of the right sizes and cannot be improved.

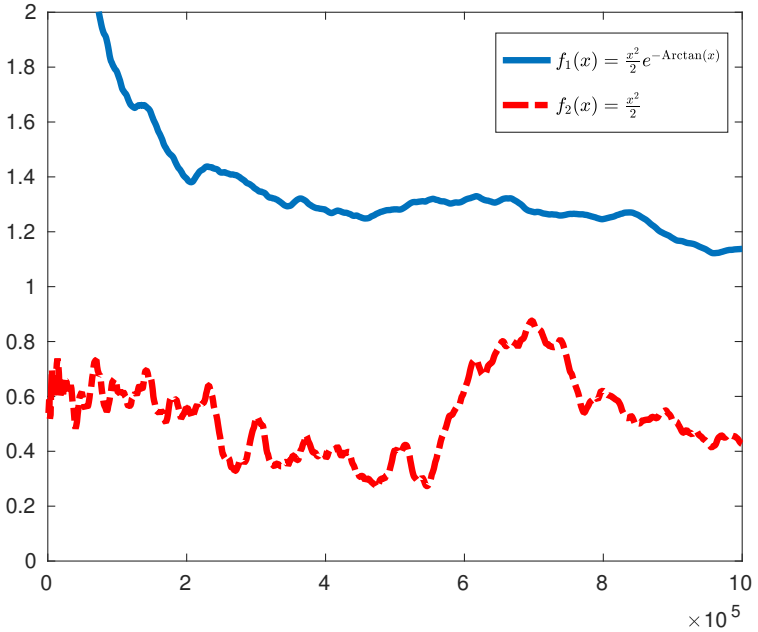

FIGURE 1. $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n}\right)$. Blue curve: $\rho = \frac{5}{4}$ and $\beta = \frac{3}{4}$ for a non locally symmetric function f_1. Red curve: $\rho = \frac{4}{3}$ and $\beta = \frac{2}{3}$ for a locally symmetric function f_2.

4. Proof of the $(L^p,\sqrt{\gamma_n})$-consistency - (Theorem 4) The main objective of this section is to prove Theorem 4 *iii*). Our analysis is based on a Lyapunov-type approach with the help of $V_p : \mathbb{R}^d \to \mathbb{R}$ defined for a given $p \geq 1$ by:

$$V_p(x) = f^p(x) \exp(\phi(f(x)).$$

We have the following result:

Theorem 6 (Convergence rate of $(\theta_n)_{n\geq 1}$ with $\mathbf{H}_\phi$) *Let $p \geq 1$ and assume $(\mathbf{H}_\phi)$ and $(\mathbf{H}^\phi_{\mathbf{\Sigma_p}})$. Let $(\gamma_n)_{n\to}$ be a non-increasing sequence such that $\gamma_n \to 0$ as $n \to +\infty$. Then,*
i) An integer $n_0 \in \mathbb{N}$ and some positive c_1 and c_2 exist such that

$$\forall n \geq n_0, \quad \mathbb{E}[V_p(\theta_{n+1})] \leq (1 - c_1\gamma_{n+1})\mathbb{E}[V_p(\theta_n)] + c_2\gamma_{n+1}^{p+1}. \tag{34}$$

ii) Furthermore, if $\gamma_n - \gamma_{n+1} = o(\gamma_{n+1}^2)$ as $n \to +\infty$, then

$$\forall n \geq 1 \qquad \mathbb{E}\left[V_p(\theta_n)\right] \leq C_p\{\gamma_n\}^p.$$

In particular,

$$\forall n \geq 1 \qquad \mathbb{E}[f^p(\theta_n)] \leq C_p\{\gamma_n\}^p.$$

Note that the condition $\gamma_n - \gamma_{n+1} = o(\gamma_{n+1}^2)$ is satisfied when $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0,1)$. Therefore, Theorem 4 *iii*) holds true.

To prove Theorem 6 *i*), we need some technical results related to ϕ and V_p. The first result is a simple sub-additive property on ϕ that essentially relies on the concavity property on $[x_0, +\infty)$.

LEMMA 1. *Assume that ϕ satisfies $(\mathbf{H}_\phi)(i)$, then a constant c_ϕ exists such that for all $x, y \in \mathbb{R}_+$:*

$$\phi(x+y) \leq \phi(x) + \phi(y) + c_\phi.$$

Proof: Since $\phi'' \leq 0$ on $[x_0, +\infty)$, the function ϕ is concave on $[x_0, +\infty)$. Hence, the function $x \mapsto \phi(x+y) - \phi(x)$ is decreasing on $[x_0, +\infty)$ and we deduce that:

$$\forall x \geq x_0 \quad \phi(x+y) \leq \phi(x) + \phi(x_0+y) - \phi(x_0).$$

Since ϕ' is decreasing on $[x_0, +\infty)$, then ϕ' is upper-bounded and a constant $C > 0$ exists such that $\phi(y + x_0) \leq \phi(y) + Cx_0$. We then deduce that:

$$\forall x \geq x_0 \quad \forall y \geq 0 \qquad \phi(x+y) \leq \phi(x) + \phi(y) + Cx_0 - \phi(x_0). \tag{35}$$

In the other situation when $x \leq x_0$, the fact that ϕ is non-decreasing yields and Equation (35) applied at point x_0 yields:

$$\phi(x+y) \leq \phi(x_0+y) \leq \phi(y) + Cx_0 \leq \phi(x) + \phi(y) + Cx_0.$$

We then obtain the desired inequality for any value of x and y in $\mathbb{R}_+$. □

The second key element of our study is a straightforward computation of the first and second derivatives of V_p.

LEMMA 2. *For any $p \in \mathbb{N}^\star$ and any $x \in \mathbb{R}^d \setminus \{\theta^\star\}$, we have:*
i)

$$\nabla V_p(x) = V_p(x)\left(p\frac{\nabla f(x)}{f(x)} + \phi'(f(x))\nabla f(x)\right).$$

ii)
$$D^2V_p(x) = V_p(x)\left[\psi_1(x)\nabla f(x)\otimes\nabla f(x) + \psi_2(x)D^2f(x)\right],$$
where ψ_1 *and* ψ_2 *are given by:*
$$\psi_1(x) := \left(\frac{p}{f(x)} + \phi'(f(x))\right)^2 - \frac{p}{f^2(x)} + \phi''(f(x)) \quad \textit{and} \quad \psi_2(x) := \frac{p}{f(x)} + \phi'(f(x)).$$

LEMMA 3. *Assume that* f *satisfies* $(\mathbf{H}_\phi)$*, then one has*
i) A constant $\alpha > 0$ *exists such that:*
$$\inf_{x\in\mathbb{R}^d} \frac{\langle \nabla V_p(x), \nabla f(x)\rangle}{V_p(x)} \geq \alpha > 0.$$
ii) For any matrix norm $\|.\|$*, a positive constant* $C > 0$ *exists such that for any* $\xi \in \mathbb{R}^d$*,*
$$\|D^2V_p(\xi)\| \leq C\left(V_{p-1}(\xi) + \frac{V_p(\xi)}{1+|\nabla f(\xi)|^2}\right).$$

Proof: Below, C refers to a large enough constant independent of ξ whose value may change from line to line.
i) We apply Lemma 2 *i)* and obtain that:
$$\forall x \in \mathbb{R}^d \setminus \{\theta^\star\} \qquad \frac{\langle \nabla V_p(x), \nabla f(x)\rangle}{V_p(x)} = p\frac{\|\nabla f(x)\|^2}{f(x)} + \phi'(f(x))\|\nabla f(x)\|^2.$$
The result then follows from Assumption $(\mathbf{H}_\phi)ii)$ and a continuity argument around $\theta^\star$.
ii) We apply Lemma 2 *ii)* and write that $\forall y \in \mathbb{R}^d$:
$$\begin{aligned}\frac{\langle y, D^2V_p(\xi)y\rangle}{\|y\|^2} &= V_p(\xi)\left[\psi_1(\xi)\langle y, \nabla f(\xi)\otimes\nabla f(\xi)y\rangle + \psi_2(\xi)\langle y, D^2f(\xi)y\rangle\right] \\ &\leq V_p(\xi)\left(\left[\frac{2p^2}{f^2(\xi)} + 2\{\phi'(f(\xi))\}^2 - \frac{p}{f^2(\xi)} + \phi''(f(\xi))\right]\|\nabla f(\xi)\|^2\right. \\ &\quad \left. + \left[\frac{p}{f(\xi)} + \phi'(f(\xi))\right]\|D^2f(\xi)\|\right).\end{aligned}$$
We now apply assumption $\mathbf{H}_\phi$: a large enough constant C exists such that:
$$\frac{\|\nabla f(\xi)\|^2}{f^2(\xi)} \leq \frac{C}{f(\xi)} \qquad \text{and} \qquad \phi'(f(\xi))^2\|\nabla f(\xi)\|^2 \leq C\phi'(f(\xi)).$$

Since $\xi \longmapsto \|D^2f(\xi)\|$ is bounded under Assumption $\mathbf{H}_\phi$ from the norm equivalence in any finite dimensional real vector space, we then have that:
$$\begin{aligned}\frac{\langle y, D^2V_p(\xi)y\rangle}{\|y\|^2} &\leq CV_p(\xi)\left[\frac{1}{f(\xi)} + \phi'(f(\xi)) + \phi''(f(\xi))\right] \\ &\leq CV_{p-1}(\xi) + \frac{CV_p(\xi)}{1+\|\nabla f(\xi)\|^2}\left(1+\|\nabla f(\xi)\|^2\right)\left(\phi'(f(\xi)) + \phi''(f(\xi))\right).\end{aligned}$$
Since $\phi''(u)$ is negative for u large enough, that ϕ' is bounded (it is a non-increasing function on $[x_0, +\infty)$) and that Assumption $\mathbf{H}_\phi$ implies that $\lim \phi'(f(\xi))|\nabla f(\xi)|^2 < +\infty$, we then deduce that:
$$\sup_{\xi\in\mathbb{R}^d}(\phi'(f(\xi)) + \phi''(f(\xi))(1+|\nabla f(\xi)|^2) < +\infty.$$
Hence,
$$\forall y \in \mathbb{R}^d \qquad \frac{\langle y, D^2V_p(\xi)y\rangle}{\|y\|^2} \leq C\left(V_{p-1}(\xi) + \frac{V_p(\xi)}{1+\|\nabla f(\xi)\|^2}\right).$$
The second assertion follows. □

The next lemma will be useful to produce an efficient descent inequality.

LEMMA 4. *Suppose that* $\mathbf{H}_\phi$ *holds and consider* $\rho \in [0,1]$. *For any* $\gamma > 0$, $\varepsilon > 0$ *define* $\xi_{\gamma,\varepsilon,x} = x + \rho\gamma\left(-\nabla f(x) + \varepsilon\right)$. *Then,*

i) A $\gamma_0 > 0$, *a constant* $C > 0$ *independent of* ρ *and* $\varepsilon > 0$ *exist such that for any* $\gamma \in [0, \gamma_0]$ *such that:*

$$f(\xi_{\gamma,\varepsilon,x}) \leq f(x) + C\gamma|\varepsilon|^2.$$

ii) If $2\gamma\|D^2 f\|_\infty \leq 1$, *then* $\forall \rho > 0 : \exists\, c_\rho > 0 : \forall x \in \mathbb{R}^d$:

$$\begin{aligned}&\gamma^2 D^2 V_p(\xi_{\gamma,\varepsilon,x})\left(-\nabla f(x) + \varepsilon\right)^{\otimes 2} \\ &\leq C(1 + |\varepsilon|^{2(p+1)}) \exp(\phi(\gamma|\varepsilon|^2)) \left(\rho\gamma V_p(x) + \gamma^2 V_p(x) + (c_\rho + 1)\gamma^{p+1}\right).\end{aligned}$$

Proof: Below, C is a positive constant whose value may change from line to line.
i) Using the Taylor formula, a $\tilde{\xi}$ exists on the segment $[x, \xi_{\gamma,\varepsilon,x}]$ such that:

$$f(\xi_{\gamma,\varepsilon,x}) = f(x) - \rho\gamma\|\nabla f(x)\|^2 + \rho\gamma\langle\nabla f(x), \varepsilon\rangle + \frac{\rho^2\gamma^2}{2} D^2 f(\tilde{\xi})\left(-\nabla f(x) + \varepsilon\right)^{\otimes 2}.$$

$\mathbf{H}_\phi$ implies that $D^2 f$ is upper bounded and $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ yields:

$$D^2 f(\tilde{\xi})\left((-\nabla f(x) + \varepsilon\right)^{\otimes 2} \leq 2\|D^2 f\|_\infty \left(\|\nabla f(x)\|^2 + \|\varepsilon\|^2\right).$$

By the elementary inequality $|\langle u, v\rangle| \leq \frac{1}{2}(\|u\|^2 + \|v\|^2)$ we deduce that:

$$\begin{aligned}f(\xi_{\gamma,\varepsilon,x}) &\leq f(x) - \rho\gamma\|\nabla f(x)\|^2 + \rho\gamma\langle\nabla f(x), \varepsilon\rangle + C\frac{\rho^2\gamma^2}{2}\left(\|\nabla f(x)\|^2 + \|\varepsilon\|^2\right) \\ &\leq f(x) + \rho\gamma\left[\frac{-1}{2} + \rho\gamma\|D^2 f\|_\infty\right]\|\nabla f(x)\|^2 + \left[\frac{\rho\gamma}{2} + \|D^2 f\|_\infty \rho^2\gamma^2\right]\|\varepsilon\|^2 \\ &\leq f(x) + \rho\gamma\|\varepsilon\|^2 \leq f(x) + \gamma\|\varepsilon\|^2,\end{aligned}$$

where in the last line we use that $\rho \leq 1$ and the condition $\gamma\|D^2 f\|_\infty \leq 1/2$. The result follows by choosing $\gamma_0 \leq C^{-1}$.

ii) We divide the proof into 4 steps.
• Step 1: Comparison between $V_r(\xi_{\gamma,\varepsilon,x})$ and $V_r(x)$. Let $r \geq 0$. Since ϕ is non-decreasing, one first deduces from *i)* that a constant $C > 0$ exists such that:

$$V_r(\xi_{\gamma,\varepsilon,x}) \leq \left(f(x) + C\gamma\|\varepsilon\|^2\right)^r \exp\left(\phi(f(x) + \gamma\|\varepsilon\|^2)\right).$$

The sub-additivity property of Lemma 1 associated with $(|a| + |b|)^r \leq 2^r(|a|^r + |b|^r)$ yields:

$$V_r(\xi_{\gamma,\varepsilon,x}) \leq 2^r\left(f^r(x) + (C\gamma)^r\|\varepsilon\|^{2r}\right) e^{\phi(f(x)) + \phi(\gamma\|\varepsilon\|^2) + c_\phi}.$$

Setting $T_{\varepsilon,\gamma,r} = (1 + \|\varepsilon\|^{2r})\exp(\phi(\gamma\|\varepsilon\|^2)$, and using that $V_0 = e^{\phi(f)}$:

$$\begin{aligned}\forall r \geq 0 \quad \exists C_r > 0 \qquad V_r(\xi_{\gamma,\varepsilon,x}) &\leq C_r \exp(\phi(\gamma\|\varepsilon\|^2)\left[V_r(x) + \gamma^r\|\varepsilon\|^{2r} V_0(x)\right] \\ &\leq C_r \exp(\phi(\gamma\|\varepsilon\|^2)\left[(1 + \|\varepsilon\|^{2r})V_r(x) + \gamma^r\|\varepsilon\|^{2r}\right] \\ &\leq C_r T_{\varepsilon,\gamma,r}\left[V_r(x) + \gamma^r\right]. \qquad (36)\end{aligned}$$

where in the second line, we used that $V_0 \leq c(1 + V_r)$.
• Step 2: Upper bound of $D^2 V_p(\xi_{\gamma,\varepsilon,x})\|.\|\nabla f(x)\|^2$. We apply Lemma 3 *ii)* with $\xi = \xi_{\gamma,\varepsilon,x}$ and we obtain that:

$$\begin{aligned}\|D^2 V_p(\xi_{\gamma,\varepsilon,x})\|.\|\nabla f(x)\|^2 &\leq C\left(V_{p-1}(\xi_{\gamma,\varepsilon,x}) + \frac{V_p(\xi_{\gamma,\varepsilon,x})}{1 + \|\nabla f(\xi_{\gamma,\varepsilon,x})\|^2}\right)\|\nabla f(x)\|^2 \\ &\lesssim \left(T_{\epsilon,\gamma,p-1}[V_{p-1}(x) + \gamma^{p-1}] + \frac{T_{\varepsilon,\gamma,p}[V_p(x) + \gamma^p]}{1 + \|\nabla f(\xi_{\gamma,\varepsilon,x})\|^2}\right)\|\nabla f(x)\|^2 \\ &\lesssim T_{\epsilon,\gamma,p-1} V_{p-1}(x)[\|\nabla f(x)\|^2 + \gamma^{p-1}\|\nabla f(x)\|^2] \\ &\quad + T_{\varepsilon,\gamma,p}\frac{\|\nabla f(x)\|^2}{1 + \|\nabla f(\xi_{\gamma,\varepsilon,x})\|^2}[V_p(x) + \gamma^p].\end{aligned}$$

Under Assumption $(\mathbf{H}_\phi)$, $V_{p-1}(x)\|\nabla f(x)\|^2 \leq CV_p(x)$ and $\gamma^{p-1}\|\nabla f(x)\|^2 \leq C\gamma^{p-1}f(x) \leq C\gamma^{p-1}(1+V_p(x))$. From the boundedness of γ and the trivial inequality since $T_{\epsilon,\gamma,p-1} \leq 2T_{\epsilon,\gamma,p}$, we then deduce that

$$\begin{aligned}\|D^2V_p(\xi_{\gamma,\varepsilon,x})\|.\|\nabla f(x)\|^2 &\lesssim T_{\epsilon,\gamma,p-1}[V_p(x)+\gamma^{p-1}] + \frac{T_{\epsilon,\gamma,p}[V_p(x)+\gamma^p]\|\nabla f(x)\|^2}{1+\|\nabla f(\xi_{\gamma,\varepsilon,x})\|^2} \\ &\lesssim T_{\epsilon,\gamma,p}\left[[V_p(x)+\gamma^{p-1}] + \frac{[V_p(x)+\gamma^p]\|\nabla f(x)\|^2}{1+\|\nabla f(\xi_{\gamma,\varepsilon,x})\|^2}\right], \end{aligned} \tag{37}$$

and we are forced to produce an upper bound of $\frac{\|\nabla f(x)\|^2}{1+|\nabla f(\xi_{\gamma,\varepsilon,x})|^2}$. According to the Taylor formula, a ξ' exists in $[x,\xi_{\gamma,\varepsilon,x}]$ such that:

$$\nabla f(x) = \nabla f(\xi_{\gamma,\varepsilon,x}) - \rho\gamma D^2 f(\xi')\left(-\nabla f(x)+\varepsilon\right),$$

and the triangle inequality yields:

$$\|\nabla f(x)\| \leq \|\nabla f(\xi_{\gamma,\varepsilon,x})\| + \|D^2 f\|_\infty \gamma(\|\nabla f(x)\| + \|\varepsilon\|),$$

so that:

$$\|\nabla f(x)\| \leq (1-\|D^2 f\|_\infty\gamma)^{-1}\left(\|\nabla f(\xi_{\gamma,\varepsilon,x})\| + \|\varepsilon\|\right).$$

The elementary inequality $(u+v)^2 \leq 2(u^2+v^2)$ leads to:

$$\|\nabla f(x)\|^2 \leq 8\|\nabla f(\xi_{\gamma,\varepsilon,x})\|^2 + \|\varepsilon\|^2).$$

As a consequence, for a large enough constant C, we have that:

$$\left(\frac{\|\nabla f(x)\|^2}{1+\|\nabla f(\xi)\|^2} + |\varepsilon|^2\right) \leq C(1+\|\varepsilon\|^2).$$

Plugging this inequality in (37) yields:

$$\|D^2V_p(\xi_{\gamma,\varepsilon,x})\|.\|\nabla f(x)\|^2 \lesssim T_{\epsilon,\gamma,p}\left(\gamma^{p-1} + [V_p(x)+\gamma^p](1+\|\varepsilon\|^2)\right),$$

and since $T_{\varepsilon,\gamma,p}(1+\|\varepsilon\|^2) \leq 3T_{\varepsilon,\gamma,p+1}$, we then conclude that:

$$\|D^2V_p(\xi_{\gamma,\varepsilon,x})\|.\|\nabla f(x)\|^2 \lesssim T_{\epsilon,\gamma,p+1}\left(\gamma^{p-1} + V_p(x)\right), \tag{38}$$

• Step 3: Upper bound of $D^2V_p(\xi_{\gamma,\varepsilon,x})\|.\|\epsilon\|^2$. We focus on the noise part ε. Using (36) and Lemma 3 *ii)* once again, we have that:

$$\|D^2V_p(\xi_{\gamma,\varepsilon,x})\|.\|\varepsilon\|^2 \lesssim T_{\varepsilon,\gamma,p+1}\left(V_{p-1}(x) + V_p(x) + \gamma^{p-1}\right). \tag{39}$$

• Step 4: Upper bound of $D^2V_p(\xi_{\gamma,\varepsilon,x})\|(-\nabla f(x)+\varepsilon)^{\otimes 2}$. We use Equations (38) and (39) and obtain:

$$\gamma^2 D^2V_p(\xi_{\gamma,\varepsilon,x})(-\nabla f(x)+\varepsilon)^{\otimes 2} \leq CT_{\varepsilon,\gamma,p+1}(\gamma^2V_{p-1}(x) + \gamma^2V_p(x) + \gamma^{p+1}).$$

To obtain the result, it is now enough to prove for any $\rho>0$, a constant c_ρ exists such that:

$$\gamma^2V_{p-1}(x) \leq \rho\gamma V^p(x) + c_\rho\gamma^{p+1}.$$

To derive this key comparison, we use the Young inequality $uv \leq \frac{u^{\bar{p}}}{\bar{p}} + \frac{v^{\bar{q}}}{\bar{q}}$ when $1/\bar{p} + 1/\bar{q} = 1$. In particular, we choose $u = \tilde{\rho}\gamma^{\frac{p-1}{p}} V^{p-1}(x)$, $v = \gamma^{1+1/p}\tilde{\rho}^{-1}$, $\bar{p} = p/(p-1)$, $\bar{q} = p$ and obtain that

$$\begin{aligned}
\gamma^2 V_{p-1}(x) &= \exp(\phi(\gamma\|\epsilon\|^2))\gamma^2 f^p(x) \\
&\leq \exp(\phi(\gamma\|\epsilon\|^2)) \left[\frac{p-1}{p}\left(\tilde{\rho}\gamma^{(p-1)/p} f^{p-1}(x)\right)^{p/(p-1)} + \frac{\gamma^{p+1}}{p\tilde{\rho}^p}\right] \\
&\leq \frac{p-1}{p}\tilde{\rho}^{p/(p-1)}\gamma V_p(x) + p^{-1}\tilde{\rho}^{-p}\gamma^{p+1}\exp(\phi(\gamma\|\epsilon\|^2)) \\
&\leq \frac{p-1}{p}\tilde{\rho}^{p/(p-1)}\gamma V_p(x) + p^{-1}\tilde{\rho}^{-p}\gamma^{p+1} V_0(x)
\end{aligned}$$

Using $V_0 \leq C(1+V_p)$ once again, we then deduce that for any $\rho > 0$, a constant c_ρ exists such that:

$$\gamma^2 V_{p-1}(x) \leq \rho\gamma V_p(x) + c_\rho \gamma^{p+1}.$$

We obtain the final upper bound:$\forall \rho > 0\,, \exists\, c_\rho > 0\,, \forall x \in \mathbb{R}^d$:

$$\gamma^2 D^2 V_p(\xi_{\gamma,\varepsilon,x})\left(-\nabla f(x) + \varepsilon\right)^{\otimes 2} \leq C T_{\epsilon,\gamma,p+1}\left(\rho\gamma V_p(x) + \gamma^2 V_p(x) + (c_\rho + 1)\gamma^{p+1}\right).$$

□

We now focus on the proof of Theorem 6 *i*).

Proof of Theorem 6:

i) We apply the second order Taylor formula to V_p and obtain that:

$$\begin{aligned}
V_p(\theta_{n+1}) = V_p(\theta_n) &- \gamma_{n+1}\langle \nabla V_p(\theta_n), \nabla f(\theta_n)\rangle + \gamma_{n+1}\langle V_p(\theta_n), \Delta M_{n+1}\rangle \\
&+ \frac{\gamma_{n+1}^2}{2} D^2 V_p(\xi_{n+1})(-\nabla f(\theta_n) + \Delta M_{n+1})^{\otimes 2},
\end{aligned}$$

where $\xi_{n+1} = \theta_n + \rho\Delta\theta_{n+1}$, $\rho \in [0,1]$. Using Lemma 3 *i*), we obtain that a $\alpha > 0$ exists such that:

$$\forall n \in \mathbb{N}^\star \qquad V_p(\theta_n) - \gamma_{n+1}\langle \nabla V_p(\theta_n), \nabla f(\theta_n)\rangle \leq V_p(\theta_n)(1 - \alpha\gamma_{n+1}). \tag{40}$$

Moreover, we have that $\mathbb{E}[\gamma_{n+1}\langle V_p(\theta_n), \Delta M_{n+1}\rangle \,|\, \mathcal{F}_n] = 0$. Finally, Lemma 4 *ii*) shows that a constant $C > 0$ exists such that for any $\rho > 0$, for all $n \in \mathbb{N}^\star$, c_ρ exists such that:

$$\begin{aligned}
&\frac{\gamma_{n+1}^2}{2} D^2 V_p(\xi_{n+1})(-\nabla f(\theta_n) + \Delta M_{n+1})^{\otimes 2} \\
&\leq C T_{\Delta M_{n+1},\gamma_{n+1},p+1}\left(\rho\gamma_{n+1} V_p(\theta_n) + \gamma_{n+1}^2 V_p(\theta_n) + (c_\rho + 1)\{\gamma_{n+1}\}^{p+1}\right).
\end{aligned}$$

This last upper bound associated with (40) and Assumption $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$ yields:

$$\begin{aligned}
&\mathbb{E}\left[V_p(\theta_{n+1}) \,|\, \mathcal{F}_n\right] \\
&\leq (1 - \alpha\gamma_{n+1})V_p(\theta_n) + \\
&\quad C\left(\rho\gamma_{n+1} V_p(\theta_n) + \gamma_{n+1}^2 V_p(\theta_n) + (c_\rho + 1)\{\gamma_{n+1}\}^{p+1}\right)\mathbb{E}\left[T_{\Delta M_{n+1},\gamma_{n+1},p+1} \,|\, \mathcal{F}_n\right] \\
&\leq (1 - \alpha\gamma_{n+1})V_p(\theta_n) + C\Sigma_p\left(\rho\gamma_{n+1} V_p(\theta_n) + \gamma_{n+1}^2 V_p(\theta_n) + (c_\rho + 1)\{\gamma_{n+1}\}^{p+1}\right) \\
&\leq (1 - (\alpha - \rho C\Sigma_p)\gamma_{n+1} + C\Sigma_p\gamma_{n+1}^2)V_p(\theta_n) + (1 + c_\rho)C\Sigma_p\{\gamma_{n+1}\}^{p+1}.
\end{aligned}$$

We now choose ρ such that $\rho C\Sigma_p = \frac{\alpha}{2}$ and determine that two non-negative constants c_1 and c_2 exist such that $\forall n \in \mathbb{N}^\star$:

$$\mathbb{E}\left[V_p(\theta_{n+1}) \,|\, \mathcal{F}_n\right] \leq \left(1 - \frac{\alpha}{2}\gamma_{n+1} + c_1\gamma_{n+1}^2\right) V_p(\theta_n) + c_2\{\gamma_{n+1}\}^{p+1}. \tag{41}$$

Theorem 6 *i*) easily follows by taking the expectation and by using that $c_1\gamma_{n+1} \leq \alpha/4$ for n large enough.

ii) We prove by induction that a large enough $C > 0$ exists such that:

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[V_p(\theta_n)\right] \leq C\left\{\gamma_n\right\}^p. \tag{42}$$

Since $\gamma_n - \gamma_{n+1} = o(\gamma_{n+1}^2)$ as $n \to +\infty$

$$\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p \leq 1 + o(\gamma_{n+1}) \quad \text{as } n \to +\infty,$$

a sufficiently large n_1 exists such that

$$\forall n \geq n_1 \qquad 0 \leq (1 - c_1\gamma_{n+1})\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p \leq 1 - \frac{c_1}{2}\gamma_{n+1}. \tag{43}$$

We can choose C_1 large enough such that Equation (42) holds true for any $n \leq n_1$ with $C \geq C_1$. For any $n_1 \in \mathbb{N}$, the result holds for any $n \leq n_1$. Assuming that the property holds at a given rank $n \geq n_1$, we then have:

$$\begin{aligned}
\mathbb{E}[V_p(\theta_{n+1})] &\leq (1 - c_1\gamma_{n+1})\mathbb{E}[V_p(\theta_n)] + c_2\{\gamma_{n+1}\}^{p+1}. \\
&\leq (1 - c_1\gamma_{n+1})C\gamma_n^p + c_2\{\gamma_{n+1}\}^{p+1}. \\
&\leq C\{\gamma_{n+1}\}^p\left[\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p(1 - c_1\gamma_{n+1}) + \frac{c_2}{C}\gamma_{n+1}\right] \\
&\leq C\{\gamma_{n+1}\}^p\left[1 - \left(\frac{c_1}{2} - \frac{c_2}{C}\right)\gamma_{n+1}\right]
\end{aligned}$$

where we used Equation (34), the induction property (42) and Inequality (43). If we choose $C \geq C_2 = \frac{c_2}{2c_1}$, then $\mathbb{E}[V_p(\theta_n)] \leq C\gamma_n^p \Longrightarrow \mathbb{E}[V_p(\theta_{n+1})] \leq C\left\{\gamma_{n+1}\right\}^p$. This ends the proof of *ii*). □

Acknowledgments The authors gratefully acknowledge Jérôme Bolte and Gersende Fort for stimulating discussions on the Kurdyka-Łojasiewicz inequality and averaged stochastic optimization algorithms. We also warmly thank Francis Bach for several constructive comments on an earlier version of our work and pointing out some useful recent references.

References

[1] Agarwal, A., P. L. Bartlett, P. Ravikumar, M. J. Wainwright. 2012. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory* **58**(5) 3235–3249. doi:10.1109/TIT.2011.2182178.

[2] Bach, F. 2014. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.* **15** 595–627.

[3] Bach, F., E. Moulines. 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems* .

[4] Bertsekas, D. P. 1999. *Nonlinear programming*. 2nd ed. Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, MA.

[5] Bolte, J., A. Daniilidis, A. Lewis. 2006. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* **17**(4) 1205–1223. doi:10.1137/050644641. URL http://dx.doi.org/10.1137/050644641.

[6] Bolte, J., A. Daniilidis, O. Ley, L. Mazet. 2010. Characterizations of lojasiewicz inequalities: subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.* (362) 3319–3363.

[7] Bolte, J., P. Nguyen, J. Peypouquet, B. W. Suter. 2016. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program. (A), to appear* 1–37.

[8] Cabot, A., H. Engler, S. Gadat. (2009). On the long time behavior of second order differential equations with asymptotically small dissipation. *Trans. Amer. Math. Soc.* **361(11)** 5983–6017.

[9] Cardot, H., P. Cénac, A. Godichon-Baggioni. 2017. Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *Ann. Statist.* **45**(2) 591–614. doi:10.1214/16-AOS1460. URL http://dx.doi.org/10.1214/16-AOS1460.

[10] Cardot, H., P. Cenac, P.A. Zitt. 2013. Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli* **19** 18–43.

[11] Casella, G., R.L. Berger. 2001. *Statistical Inference*. Duxbury Press.

[12] Cesa-Bianchi, Nicolò, Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge University Press, Cambridge. URL https://doi.org/10.1017/CBO9780511546921.

[13] Duflo, M. 1997. *Random Iterative Models, Adaptive algorithms and stochastic approximations*. Applications of Mathematics, Springer-Verlag, New-York.

[14] Fort, G. 2015. Central limit theorems for stochastic approximation with controlled Markov chain dynamics. *ESAIM Probab. Stat.* **19** 60–80. doi:10.1051/ps/2014013. URL http://dx.doi.org/10.1051/ps/2014013.

[15] Gadat, S., L. Miclo. 2013. Spectral decompositions and l2-operator norms of toy hypocoercive semigroups. *Kinetic and Related Models* **6** 317–372.

[16] Gadat, S., F. Panloup, S. Saadane. (2018). Stochastic heavy ball. *Electronic Journal of Statistics* .

[17] Kemperman, J.H.B. 1987. The median of a finite measure on a banach space. *Statistical data analysis based on the L^1-norm and related methods (Neuchtel, 1987)* 217–230.

[18] Kurdyka, K. 1998. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)* **48**(3) 769–783. URL http://www.numdam.org/item?id=AIF_1998__48_3_769_0.

[19] Lojasiewicz, S. 1963. Une propriété topologique des sous-ensembles analytiques réels. *Editions du centre National de la Recherche Scientifique, Paris, Les Équations aux Dérivées Partielles* 87–89.

[20] N. Flammarion, F. Bach. 2015. From averaging to acceleration, there is only a step-size. *Proceedings of the International Conference on Learning Theory (COLT)* .

[21] Nemirovski, A., D. Yudin. 1983. Problem complexity and method efficiency in optimization. *Wiley-Interscience Series in Discrete Mathematics*. **John Wiley, XV**.

[22] Nesterov., Y. 1983. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady* **27(2)** 372–376.

[23] Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization. A basic course.*. Applied Optimization, Kluwer Academic Publishers, Boston, MA.

[24] Polyak, B. T., A. Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* **30** 838–855.

[25] Robbins, H., S. Monro. 1951. A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

[26] Ruppert, D. 1988. Efficient estimations from a slowly convergent robbins-monro process. *Technical Report, 781, Cornell University Operations Research and Industrial Engineering* .

[27] Villani., C. 2009. Hypocoercivity. *Mem. Amer. Math. Soc.* **202(950)**.

5. Technical lemmas for Theorem 2 The next lemma is important to obtain the stability of the change of basis from one iteration to another in our spectral analysis of $(\hat{\theta}_n)_{n\geq 1}$.

LEMMA 5. *Assume that $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0,1)$. Let $\mu > 0$. Then, a constant C and an integer n_0 exist such that*

$$\forall n \geq n_0, \qquad |\epsilon_{\mu,n} - \epsilon_{\mu,n+1}| \leq C n^{\beta-2}$$

Proof: We choose n_0 such that $1-\mu\gamma_n n<0$ for all $n\geq n_0$. Then, the desired inequality comes from a direct computation:

$$\begin{aligned}\epsilon_{\mu,n}-\epsilon_{\mu,n+1} &= \frac{1-\mu\gamma_n}{1-\mu\gamma_n n}-\frac{1-\mu\gamma_{n+1}}{1-\mu\gamma_{n+1}(n+1)}\\ &= \frac{(1-\mu\gamma_n)(1-\mu\gamma_{n+1}(n+1))-(1-\mu\gamma_{n+1})(1-\mu\gamma_n n)}{(1-\mu\gamma_n n)(1-\mu\gamma_{n+1}(n+1))}\\ &= \mu\frac{(\gamma_{n+1}-\gamma_n)+(n\gamma_n-(n+1)\gamma_{n+1})+\mu\gamma_n\gamma_{n+1}}{(1-\mu\gamma_n n)(1-\mu\gamma_{n+1}(n+1))}\end{aligned}$$

Now, if C denotes a constant that only depends on μ and β (whose value may change from line to line), we then have the following inequalities:

$$|\gamma_{n+1}-\gamma_n|\leq Cn^{-(1+\beta)},\ |n\gamma_n-(n+1)\gamma_{n+1}|\leq Cn^{-\beta} \text{ and } \gamma_n\gamma_{n+1}\leq Cn^{-2\beta}.$$

Since $\beta<1$, the denominator is equivalent to $n^{2-2\beta}$ and we obtain that

$$|\epsilon_{\mu,n}-\epsilon_{\mu,n+1}|\leq C\frac{n^{-\beta}}{n^{2-2\beta}}=Cn^{\beta-2}, \tag{44}$$

which ends the proof. □

LEMMA 6. *Under the assumptions of Proposition 3.4, we have:*
i) For any $i\in\{1,\dots,d\}$, $\omega_n(i)=\mathbb{E}[(\widetilde{Z}_n^{(1)})_i(\widetilde{Z}_n^{(2)})_i]$ *satisfies* $\forall n\geq n_0$,

$$\omega_{n+1}(i)=(1-\gamma_{n+1}\mu_i^\star)\frac{n}{n+1}\omega_n(i)+O\left(\frac{\gamma_{n+1}}{n+1}\right)+O(\gamma_{n+1}\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|]).$$

ii) The following recursion holds for any $n\geq n_0$,

$$\begin{aligned}\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2]=\left(1-\frac{1}{n+1}\right)^2\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2]+\sum_{i=1}^d\alpha_n^i\omega_n(i)+\frac{\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|]}{n}\\ +\frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2}+O\left(\frac{\sqrt{\gamma_n}}{n^2}\vee\frac{1}{n^3\gamma_n}\right),\end{aligned}$$

where α_n^i *is defined in (31) and satisfies* $|\alpha_n^i|\lesssim\gamma_n^{-1}n^{-2}$, $i=1,\dots,d$.

Proof: Set $\underline{\mu}=\min\{\mu_i^\star,i=1,\dots,d\}>0$. Recall that $n_0\in\mathbb{N}$ is such that $1-\underline{\mu}\gamma_n n<0$ for all $n\geq n_0$. For all $n\geq n_0$, Υ_n and Ω_n are well-defined deterministic matrices and since for a given $\mu>0$, $\epsilon_{\mu,n}\sim(n\gamma_n)^{-1}$ and $\epsilon_{\mu,n}-\epsilon_{\mu,n+1}=O(n^{-2}\gamma_n^{-1})$ (see (44)), we have

$$\gamma_{n+1}\|\Upsilon_n\|=O\left(\frac{1}{n}\right)\quad\text{and}\quad\gamma_{n+1}\|\Omega_n\|=O\left(\frac{1}{n^2}\right). \tag{45}$$

Now, let us prove the first statement.
i) Using (29), we have

$$\begin{aligned}\omega_{n+1}(i)=(1-\gamma_{n+1}\mu_i^\star)\left(1-\frac{1}{n+1}\right)\omega_n(i)+O(\gamma_{n+1}\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|])\\ +\gamma_{n+1}^2\mathbb{E}[\{Q\Delta M_{n+1}\}_i\{\Upsilon_n Q\Delta M_{n+1}\}_i]+O(\gamma_{n+1}r_n^{(1)})\end{aligned}$$

where

$$r_n^{(1)}=\|\Omega_n\|\left(\frac{\mathbb{E}|\widetilde{Z}_n^{(1)}|^2}{\gamma_{n+1}}+\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(1)}|]\right)+\|\Upsilon_n\|\left(\mathbb{E}[|\widetilde{Z}_n^{(1)}|.|\theta_n|^2]+\gamma_{n+1}\mathbb{E}|\theta_n|^4\right).$$

The Cauchy-Schwarz inequality, the fact that $|\widetilde{Z}_n^{(1)}| = |\theta_n|$ and the consistency condition lead to

$$\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(1)}|] \leq \left\{\mathbb{E}[|\theta_n|^4|\}\right\}^{1/2}\left\{\mathbb{E}[|\widetilde{Z}_n^{(1)}|^2]\right\}^{1/2} \leq \gamma_{n+1}^{3/2}.$$

Therefore, (45) yields:

$$\gamma_{n+1} r_n^{(1)} \lesssim \frac{1}{n^2}\left(1+\gamma_{n+1}^{\frac{3}{2}}\right) + \frac{1}{n}\left(\gamma_{n+1}^{\frac{3}{2}} + \gamma_{n+1}^3\right) = o\left(\frac{\gamma_n}{n}\right).$$

In the meantime, under $(\mathbf{H_S})$ and because $Q \in O_d(\mathbb{R})$, we have:

$$\begin{aligned}\forall i \in \{1,\dots,d\} \qquad |\mathbb{E}[\{Q\Delta M_{n+1}\}_i \{\Upsilon_n Q \Delta M_{n+1}\}_i]| &\lesssim \|\Upsilon_n\|\mathbb{E}[|\Delta M_{n+1}|^2] \\ &\lesssim \|\Upsilon_n\|\mathbb{E}\|S(\theta_n)\| \\ &\lesssim \|\Upsilon_n\|(1+\mathbb{E}|\theta_n|) \\ &\lesssim \|\Upsilon_n\|.\end{aligned}$$

We therefore deduce from (45) and from the previous lines that

$$\forall i \in \{1,\dots,d\} \qquad \gamma_{n+1}^2 |\mathbb{E}[\{Q\Delta M_{n+1}\}_i \{\Upsilon_n Q \Delta M_{n+1}\}_i]| \lesssim \frac{\gamma_n}{n}.$$

ii) We define $\Delta N_{n+1} = \Upsilon_n Q \Delta M_{n+1}$ and recall that α_n^i is defined in (31) by $\alpha_n^i = 2(1-(n+1)^{-1})(\Omega_n)_{i,i}$. Starting from (29) and $|\widetilde{Z}_n^{(1)}| = |\theta_n|$, we use that Ω_n is a diagonal matrix so that

$$\begin{aligned}\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] = \left(1-\frac{1}{n+1}\right)^2 \mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + \sum_{i=1}^d \alpha_n^i \omega_n(i) + \gamma_{n+1}^2 \mathbb{E}|\Delta N_{n+1}|^2 \\ + O\left(\gamma_{n+1}\|\Upsilon_n\|\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|]\right) + O(\gamma_{n+1} r_n^{(2)}),\end{aligned}$$

where $r_n^{(2)}$ is defined by

$$r_n^{(2)} = \frac{\|\Omega_n\|^2 \mathbb{E}|\theta_n|^2}{\gamma_{n+1}} + \|\Omega_n\|\|\Upsilon_n\|\mathbb{E}|\theta_n|^3 + \gamma_{n+1}\|\Upsilon_n\|^2\mathbb{E}|\theta_n|^4.$$

The $(L^4, \sqrt{\gamma_n})$-consistency, the Jensen inequality and (45) yield

$$\gamma_{n+1} r_n^{(2)} = O\left(\frac{1}{\gamma_n n^4} + \frac{\sqrt{\gamma_n}}{n^3} + \frac{\gamma_n^2}{n^2}\right) = O\left(\frac{1}{n^3}\right)$$

since $\gamma_n \leq c n^{-\frac{1}{2}}$. To achieve the proof, it remains to show that

$$\mathbb{E}|\Delta N_{n+1}|^2 = \frac{\mathrm{Tr}(\Sigma^\star)}{n^2} + O\left(\frac{\sqrt{\gamma_n}}{n^2} \vee \frac{1}{n^3\gamma_n}\right) \tag{46}$$

First, set $B_n = Q^T \Upsilon_n^2 Q$. Using that Υ_n is a diagonal matrix, we have

$$\begin{aligned}|\Delta N_{n+1}|^2 = \mathrm{Tr}(|\Delta N_{n+1}|^2) &= \mathrm{Tr}(\Delta N_{n+1}^T \Delta N_{n+1}) \\ &= \mathrm{Tr}(\Delta M_{n+1}^T B_n \Delta M_{n+1}) \\ &= \mathrm{Tr}(B_n \Delta M_{n+1} \Delta M_{n+1}^T)\end{aligned}$$

Since the trace is a linear application and B_n is a deterministic matrix,

$$\mathbb{E}[|\Delta N_{n+1}\}|^2|\mathcal{F}_n] = \mathrm{Tr}(B_n \mathbb{E}[\Delta M_{n+1}\Delta M_{n+1}^T|\mathcal{F}_n]) = \mathrm{Tr}(B_n S(\theta_n)) \tag{47}$$

where we applied Assumption ($\mathbf{H_S}$). We also have $S(\theta_n) = S(\theta^\star) + O(|\theta_n|)$. For B_n, we first remark that

$$\gamma_{n+1}\Upsilon_n = (n+1)^{-1}\{D^\star\}^{-1} + \Delta_{n+1}$$

where $(\Delta_n)_{n\geq 0}$ is a sequence of matrices defined by:

$$\Delta_n = \mathrm{Diag}\left\{\frac{1-(n+1)\{\mu_i^\star\}^2\gamma_{n+1}^2}{(n+1)\mu_i^\star((n+1)\gamma_{n+1}\mu_i^\star - 1)} + \frac{\gamma_{n+1}}{n+1},\, i=1,\dots,d\right\}.$$

Using that $n\gamma_n^2 \to 0$ as $n \to +\infty$, one easily checks that

$$\|\Delta_n\| \lesssim \frac{1}{n^2\gamma_n} + \frac{\gamma_n}{n} \lesssim \frac{1}{n^2\gamma_n}.$$

As a consequence,

$$\begin{aligned}\gamma_{n+1}^2 B_n &= Q^T\{\gamma_{n+1}\Upsilon_n\}^2 Q \\ &= Q^T\{(n+1)^{-1}D^\star + \Delta_{n+1}\}^2 Q \\ &= (n+1)^{-2}Q^T\{D^\star\}^{-2}Q + O\left(\frac{1}{n^3\gamma_n}\right).\end{aligned}$$

It follows from (47) that

$$\begin{aligned}\gamma_{n+1}^2\mathbb{E}[|\Delta N_{n+1}\}|^2|\mathcal{F}_n] &= \gamma_{n+1}^2\mathrm{Tr}(B_n\Delta M_{n+1}\Delta M_{n+1}^T) \\ &= \frac{\mathrm{Tr}(\{\Lambda^\star\}^{-2}S(\theta^\star))}{(n+1)^2} + O\left(\frac{\mathbb{E}|\theta_n|}{n^2} \vee \frac{1}{n^3\gamma_n}\right) \\ &= \frac{\mathrm{Tr}(\{\Lambda^\star\}^{-2}S(\theta^\star))}{(n+1)^2} + O\left(\frac{\sqrt{\gamma_n}}{n^2} \vee \frac{1}{n^3\gamma_n}\right)\end{aligned}$$

because which leads to (46) and achieves the proof. □

LEMMA 7. *Assume that $(u_n)_{n\geq 0}$ is a sequence which satisfies for all $n \geq n_0$ and for a given $\mu > 0$:*

$$u_{n+1} = (1-\gamma_{n+1}\mu)\frac{n}{n+1}u_n + \beta_{n+1}$$

with $\beta_n \lesssim \gamma_n n^{-1}$. Then, $u_n = O(n^{-1})$.

Proof: With the convention $\prod_\emptyset = 1$ and $\sum_\emptyset = 0$, we have for every $n \geq n_0$:

$$u_n = \left(\prod_{k=n_0+1}^n (1-\gamma_k\mu)\frac{k}{k+1}\right)u_{n_0} + \sum_{k=n_0+1}^n \beta_k \prod_{\ell=k+1}^n (1-\gamma_\ell\mu)\frac{\ell}{\ell+1}.$$

Using that for any $x > -1$, $\log(1+x) \leq x$, we obtain for every $n \geq n_0+1$

$$\prod_{k=n_0+1}^n (1-\gamma_k\mu)\frac{k}{k+1} \leq \frac{n_0}{n+1}e^{-\mu(\Gamma_n - \Gamma_{n_0})} \leq C_{n_0}\frac{e^{-\Gamma_n}}{n+1} = O(n^{-1})$$

and,

$$\sum_{k=n_0+1}^n \beta_k \prod_{\ell=k+1}^n (1-\gamma_\ell\mu)\frac{\ell}{\ell+1} \leq \frac{1}{n+1}\left(e^{-\mu\Gamma_n}\sum_{k=n_0+1}^n \beta_k(k+1)e^{\mu\Gamma_k}\right).$$

But $\beta_k(k+1) \lesssim \gamma_{k+1}$. Thus, since $x \mapsto xe^{\mu x}$ is increasing on $\mathbb{R}_+$,

$$\sum_{k=n_0+1}^{n} \beta_k(k+1)e^{\mu\Gamma_k} \lesssim \sum_{k=n_0+1}^{n} \gamma_{k+1}e^{\mu\Gamma_k} \leq \int_{\Gamma_{n_0+1}}^{\Gamma_{n+1}} e^{\mu x}dx$$

and hence,

$$\frac{1}{n+1}\left(e^{-\mu\Gamma_n}\sum_{k=n_0+1}^{n} \beta_k(k+1)e^{\mu\Gamma_k}\right) \leq \frac{C_{n_0}}{n+1}.$$

The result follows. □

Remark 5.1 *By the expansion* $\log(1+x) = x + c(x)x^2$ *where* c *is bounded on* $[-1/2, 1/2]$*, a slight modification of the proof leads to* $\liminf_{n\to+\infty} nu_n > 0$ *when* $\sum \gamma_k^2 < +\infty$*.*

LEMMA 8. *For any sequence* $(u_n)_{n\geq 0}$ *that satisfies*

$$\forall n \geq 0 \qquad u_{n+1} \leq u_n\left(1 - \frac{1}{n+1}\right)^2(1+2n^{-r}) + \frac{V}{(n+1)^2} + \bar{c}n^{-q},$$

with $r \geq 1$ *and* $q \geq 2$*, then a large enough constant* C *independent of* n *exists such that*

$$\forall n \geq 1 \qquad u_n \leq \frac{V}{n} + Cn^{-\{r\wedge(q-1)\}}.$$

Proof: We establish the result using an induction and denote by $\alpha = r \wedge q$ The statement of the lemma is obvious for $n = 1$ by choosing a large enough C. Assuming now that the result holds for the integer n, we write

$$\begin{aligned} u_{n+1} &\leq \left(\frac{n}{n+1}\right)^2(1+2n^{-r})\left[\frac{V}{n} + Cn^{-\alpha}\right] + \frac{V}{(n+1)^2} + \bar{c}n^{-q} \\ &\leq V\left[\frac{n}{(n+1)^2} + \frac{1}{(n+1)^2}\right] + 2V\left(\frac{n}{n+1}\right)^2 n^{-(r+1)} \\ &\quad + Cn^{-\alpha}\left(\frac{n}{n+1}\right)^2 + 2Cn^{-(\alpha+r)}\left(\frac{n}{n+1}\right)^2 + \bar{c}n^{-q} \\ &= \frac{V}{n+1} + C(n+1)^{-\alpha}\mathcal{A}_n \end{aligned}$$

where

$$\mathcal{A}_n := \frac{2Vn^{1-r}(n+1)^{\alpha-2}}{C} + \left(\frac{n}{n+1}\right)^{2-\alpha} + 2\frac{n^{2-\alpha-r}}{(n+1)^{2-\alpha}} + \frac{\bar{c}}{C}n^{-q}(n+1)^{\alpha}$$

We now choose $\alpha < 2$ and use the first order approximations:

$$\mathcal{A}_n = 1 - (2-\alpha)n^{-1} + C^{-1}\left[2Vn^{\alpha-(1+r)} + 2n^{-r} + \bar{c}n^{\alpha-q}\right] + o(n^{-1}).$$

Then, a large enough C exists such that $\mathcal{A}_n \leq 1$ for any $n \geq 1$ as soon as the powers of n are lower than 1 inside the brackets on the right hand side of the equality above. Hence, α should be chosen such that $\{(1+r)-\alpha\} \wedge \{r\} \wedge \{\alpha - q\} \geq 1$ and the largest possible value of α corresponds to the choice

$$\alpha = (q-1) \wedge r.$$

For such a choice, a large enough C exists such that the recursion holds, which ends the proof of Lemma 8. □

6. Growth at infinity under the KL gradient inequality In this section, we prove the property (11) of Proposition 2.2. Without loss of generality, we can assume that $\theta^\star = f(\theta^\star) = 0$.

Proof: Consider $0 \leq t \leq s$ and $x \in \mathbb{R}^d$. We then associate the solution of the differential equation associated to the flow $-\nabla f$ initialized at x:

$$\chi_x(0) = x \qquad \text{and} \qquad \dot{\chi}_x = -\nabla f(\chi_x).$$

The length of the curve $L(\chi_x, t, s)$ is defined by

$$L(\chi_x, t, s) = \int_t^s \|\dot{\chi}_x(\tau)\| d\tau.$$

Under Assumption $(\mathbf{H^r_{KL}})$, we can consider $\varphi(a) = \frac{a^{1-r}}{1-r}$ and we have that

$$\varphi'(f(x))\|\nabla f(x)\| \geq m > 0.$$

We now observe that $e : s \longmapsto \varphi(f(\chi_x(s)))$ satisfies:

$$\begin{aligned} e'(\tau) &= \varphi'(f(\chi_x(\tau)))\langle \nabla f(\chi_x(\tau)), \dot{\chi}_x(\tau)\rangle \\ &= -\varphi'(f(\chi_x(\tau)))\|\nabla f(\chi_x(\tau))\|^2 \\ &\leq -m\|\dot{\chi}_x(\tau)\| \end{aligned}$$

We deduce that:

$$e(t) - e(s) = \int_s^t e'(\tau) d\tau \geq m \int_t^s \|\dot{\chi}_x(\tau)\| d\tau \geq m L(\chi_x, t, s) \tag{48}$$

Now choosing $t = 0$ and $s \longrightarrow +\infty$, we have $e(0) - \lim_{s\longrightarrow+\infty} e(s) = \varphi(f(x)) - \varphi(\min f) = \varphi(f(x))$, and Equation (48) yields

$$\varphi(f(x)) \geq m L(\chi_x, 0, +\infty) \geq m\|x\|$$

because $\chi_x(+\infty) = \arg\min f = 0$. We deduce that

$$f(x) \geq \varphi^{-1}(m\|x\|) = \{m(1-r)\}^{\frac{1}{1-r}} \|x\|^{\frac{1}{1-r}}.$$

which is the desired conclusion. □